Extracting Human Behaviour and Personality Traits from Social Media

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Institute of Sustainable Industries and Liveable Cities

Victoria University

By

Ravinder Singh

April 2021

Acronyms

ASB: Antisocial Behaviour

NLP: Natural Language Processing

IR: Information Retrieval

ML: Machine Learning

AI: Artificial Intelligence

LR: Logistic Regression

SVM: Support Vector Machine

RF: Random Forest

DT: Decision Tree

NB: Naïve Bayes

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

LSTM: Long Short-Term Memory Network

GRU: Gated Recurrent Units

LIWC: Linguistic Inquiry and Word Count

EXTRACTING HUMAN BEHAVIOUR AND PERSONALITY TRAITS FROM SOCIAL MEDIA

Ravinder Singh
Principle Supervisor: Prof Yanchun Zhang

Abstract

Online social media has evolved as an integral part of human society. It facilitates collaboration and information flow, and has emerged as a crucial instrument for business and government organizations alike. Online platforms are being used extensively for work, entertainment, collaboration and communication. These positive aspects are, however, overshadowed by their shortcomings. With the constant evolution and expansion of social media platforms, a significant shift has occurred in the way some humans interact with others. Online social media platforms have inadvertently emerged as networking hubs for individuals exhibiting antisocial behaviour (ASB), putting vulnerable groups of people at risk. Online ASB is one of the most common personality disorders seen on these platforms, and is challenging to address due to its complexities.

Human rights are the keystones of sturdy communities. Respect for these rights, based on the values of equality, dignity and appreciation, is vital and an integral part of strong societies. Every individual has a fundamental right to freely participate in all legal activities, including socializing in both the physical and online worlds. ASB, ranging from threatening, aggression, disregard for safety and failure to conform to lawful behaviour, deter such participation and must

be dealt with accordingly. Online ASB is the manifestation of everyday sadism and violates the elementary rights (to which all individuals are entitled) of Its victims. Not only does it interfere with social participation, it also forces individuals into anxiety, depression and suicidal ideation. The consequences of online ASB for victims' and families' mental health are often far-reaching, severe and long-lasting, and can even create a social welfare burden. The behaviour can, not only inhibit constructive user participation with social media, it defies the sole purpose of these platforms: to facilitate communication and collaboration at scale. ASB needs to be detected and curtailed, encouraging fair user participation and preventing vulnerable groups of people from falling victim to such behaviour.

Considering the large variety, high contribution speed and high volume of social media data, a manual approach to detecting and classifying online ASB is not a feasible option. Furthermore, a traditional approach based on a pre-defined lexicon and rule-based feature engineering may still fall short of capturing the subtle and latent features of the diverse and enormous volume of social media data. State-of-the-art deep learning, which is a sub-field of machine learning, has produced astonishing results in numerous text classification undertakings, and has outperformed the aforementioned techniques. However, given the complexity associated with implementing deep learning algorithms and their relatively recent development, models based on the technology have significantly been under-utilized when working with online behaviour studies.

Specifically, no prior study has undertaken the task of fine-grained and user-generated social media content classification related to online ASB utilizing the deep learning technology.

This thesis introduces a novel three-part framework, based on deep learning, with the objectives of: *i) Detecting behaviour and personality traits from online platforms; (ii) Binary detection of online antisocial behaviour and (iii) Multiclass antisocial behaviour detection from social media corpora.* A high accuracy classification model is presented proceeded by extensive experimentation with different machine learning and deep learning algorithms, fine tuning of hyper-parameters, and using different feature extraction techniques. Disparate behaviour and personality traits, including ASB and its four variants are detected with a significantly high accuracy from online social media platforms. Along the way, three medium-sized gold standard benchmark data set have been constructed. The proposed approach is seminal and offers a step towards efficient and effective methods of online ASB prevention. The approach and the findings within this thesis are significant and crucial as these lay the groundwork for detecting and eliminating all types of undesirable and unacceptable social behaviour traits from online platforms.

DOCTOR OF PHILOSOPHY DECLARATION

I, Ravinder Singh, declare that the PhD thesis entitled *Extracting Human Behaviour and Personality Traits from Social Media* is no more than 80, 000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

Signature  Date 3/04/2021

This thesis is dedicated to my parents, wife, and daughter for their love, encouragement and endless support.

ACKNOWLEDGEMENTS

I owe my heart-felt gratitude to all those who have made this thesis possible. First and foremost, I am grateful to my principal supervisor Professor Yanchun Zhang, who has been a tremendous mentor. I thank him for his patience and encouragement throughout my PhD journey. I am also grateful to my associate supervisors Professor Hua Wang and Professor Yuan Miao for their continuous guidance and support. Without all your support, this thesis would not have been possible.

I am thankful to Professor Ron Adams, Associate Professor Gavin Ivey, Professor Ronald D Francis, Dr Rose Lucas, Dr Anwaar Ul-Haq, Dr Khandakar Ahmed, and Dr Huy Quan Vu for sharing their experiences and knowledge, and helping me out when I was stuck. Your indispensable suggestions and advice got me through my PhD journey.

I am thankful to Victoria University and all its staff members with whom I have interacted with during my PhD journey. You all were very helpful and professional, and above all are awesome in everything you do to make our university lives as comfortable as you can by providing us with all the services and support. Special thanks to Professor Anne-Marie Hede, Elizabeth Smith for her continuous reminders about the PhD milestones ☺, Randall Robinson for his calm and assuring smiles, Debra Fitzpatrick for her support related to my scholarship, Cameron Barrie for library services, Palmina Fichera, Jo Xuereb,

and Meika Scholz for the administrative support. I cannot thank you all enough for making my PhD dream a reality.

I would also like to thank all my PhD colleagues Dinesh Pandey, Sarath Kumar, Saurav Dahal, Santosh Kaini, Kevin Du, Alice Kho, Ruwangi Fernando, Shekha Chenthara, Rubina Sarki, Sudha Subramani and Ye Wang for their continuous support and encouragement. There were days when I felt very low and stuck, and talking to these guys, learning from their experiences, helped me get back on track. I hope that these friendships forged at VU will last forever and we can stay in touch to inspire each other.

My PhD research journey was supported by the Research Training Program from the Australian Government's Department of Education, Skills and Employment and Victoria University. The financial support that I received to complete my thesis is gratefully acknowledged and appreciated.

I am grateful to all my friends and family members, who stood by me and encouraged me while I pursued my PhD degree. All your unconditional love and support during good and hard times is greatly appreciated. Above all, I want to thank God for all the blessings that I take for granted on a daily basis, for all the lessons (especially the hard ones) for every hardship was followed by ease, and for healing me when I was broken. You have made me who I am today. Thank you.

PUBLICATIONS

[1] Ravinder Singh, Sudha Subramani, Jiahua Du, Yanchun Zhang, Hua Wang, Khandakar Ahmed, Zhenxiang Chen. "Deep Learning for Multi-class Antisocial Behaviour Identification from Twitter," Volume:8, IEEE Access, 2020, IEEE Press.

[2] Ravinder Singh, Yanchun Zhang, Hua Wang, Yuan Miao, Khandakar Ahmed. "Deep learning for Antisocial Behaviour Analysis on Social Media". 24th International Conference Information Visualisation, IEEE 2020.

[3] Ravinder Singh, Yanchun Zhang, and Hua Wang, "Exploring Human Mobility Patterns in Melbourne Using Social Media Data," in Australasian Database Conference, 2018: Springer, pp. 328-335.

[4] Ravinder Singh, Jiahua Du, Yanchun Zhang, Hua Wang, Yuan Miao, Omid Ameri Sianaki, Anwaar Ulhaq. "A Framework for Early Detection of Antisocial Behaviour on Twitter Using Natural Language Processing," in Conference on Complex, Intelligent, and Software Intensive Systems, 2019: Springer, pp. 484-495.

[5] Ravinder Singh, Yanchun Zhang, Hua Wang, Yuan Miao, Khandakar Ahmed. "Investigation of Social Behaviour Patterns using Location-Based Data-

A Melbiourne Case Study. EAI Endorsed Transaction on Scalable Information Systems, 2020.

[6] Ravinder Singh, Yanchun Zhang, Hua Wang, Yuan Miao, and Khandakar Ahmed. "Antisocial Behaviour Analyses Using Deep Learning," in International Conference on Health Information Science, 2020: Springer, pp. 133-145.

[7] Ravinder Singh, Sudha Subramani, Jiahua Du, Yanchun Zhang, Hua Wang, Khandakar Ahmed. “Antisocial Behaviour Identification from Twitter Feeds Using Traditional Machine Learning Algorithms and Deep Learning.” PLOS ONE, 2021 (under review).

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Humans exhibit behaviour traits during interactions with their surroundings. These behaviour traits are comprised of state of mind, personality and behaviour patterns [1]. Researchers in social and behaviour science have long studied these patterns and traits using both qualitative and quantitative research methodologies. Surveys, interviews and questionnaires have typically been used to capture subtle elements related to human behaviour and personality to test hypotheses and to form new theories. Studies and experimentations related to social and ASB traits had been relatively straightforward. However, for the last couple of decades there has been a fundamental change in the way humans interact with their surrounding environment and social circles. With the advent of the Internet and social media, most of these interactions have moved online. Computers, mobile phones, smart watches and the other IoT-related gadgets have become the new norm for expressing personality and behaviour traits. Most of daily life, and the activities within it, has moved behind screens, leading to a fundamental shift in the ways humans interact and exhibit behaviour traits. The traditional ways of studying social and ASB traits are not entirely relative, and in some instances have become obsolete in capturing and studying human behaviour in the age of the Internet and social media. The problem is further exacerbated when dealing with unacceptable and ASB that can be exhibited

freely online without repercussions and implications. This can be conveniently attributed to the inherent characteristics of anonymity that have become an integral part of the Internet and social media platforms. Not only has it led to a social power shift, which has been traditionally accredited to the way humans behave, it has turned most of the established social and behaviour science theories on their head. In the real world, a person with power (social, physical or financial) can exert ASB on a subordinate, leveraging that power. However, in an online world, these powers do not play a significant role, and a person who may be a subordinate can stay anonymous and exert ASB over others, including the powerful.

Online platforms have their own advantages and benefits, however these come at a cost. ASV in its online incarnation is an offshoot of online platforms. It appears online in its many forms such as failure to conform to societal norms, threatening, disrespecting the law, intentional aggression, disregard for the safety of self and others, hostility, deceitfulness, etc. It is a common pattern of violation and disregard for the rights and well-being of others. It is a behaviour that is considered disruptive to our society and communities; whether online or offline.

The use of the term ASB in the context of online communities and in the common lexicon is relatively new. Nonetheless, it has been used in the psychological world for years and has been defined as 'an unwanted behaviour' due to a

personality disorder. Many people describe it as a behaviour that is deemed contrary to many norms that prevail in our society. Researchers, however, find it difficult to exactly define such behaviour as there could be infinite number of acts that may fall under the umbrella term ‘antisocial behaviour'. There are many factors that contribute to one's developing and exhibiting antisocial behaviour and these include genes, environmental stressors, neurobiology, maternal depression, adverse socioeconomics, pure nutrition intake, parental and societal rejection, and sociocultural factors.

Online platforms have measures in place to depict nudity and pornography, and have recently been working on detecting fake news, altered images and videos, however no attention has been paid to detecting and eliminating online ASB. Online ASB has become a societal problem, which has led to many victims experience stress, developing depression and anxiety, and in some instances committing suicide. Therefore, the issue demands serious consideration. Due to the sheer number of online users and the amount of data being produced on these platforms, ASB manually is not a practical approach. These platforms need to implement solutions that work at scale.

The motivation behind this research project is to discover a methodology that can be used at scale, not only to detect online ASB but also to identify its variants, enabling platforms to take appropriate actions to eliminate it. To tackle the stated problem, advanced deep learning, machine learning and natural

language processing techniques are required to identify critical posts accurately from the sheer amount of online data. Automating the process will enable reduction in response time and minimise the detrimental impact on the mental health of victims, making online platforms a safer place for everyone. To detect actionable knowledge from social media platforms for public health mining has already been proven successful [2-4]. To use such knowledge and techniques to detect and eliminate online ASB is, therefore, a logical next step. Deep learning, which is an advanced approach of implanting neural networks, has demonstrated encouraging results in numerous text classification studies [5] because of its inherent ability to capture subtle and latent features from a text corpus. Feature extraction techniques using word embedding in deep learning can identify semantic and syntactic relationships between phrases and terms with a significant high accuracy. Despite these abilities, deep learning techniques are considered non-intuitive, task specific, and highly empirical. The performance of a model built using machine learning in general, and deep learning in particular, is diligently associated with a case study.

The body of research on ASB, in its online incarnation, is in a nascent form with a very few studies conducted in the area. A very small number of studies have investigated the use of cutting-edge technologies to detect and eliminate unacceptable online behaviour. Deep learning and its advanced feature extraction techniques have not been utilized to study online ASB. The primary aim of this thesis is to develop a methodology and framework to automatically

detect and classify online ASB from a highly disparate and unstructured corpora of online data streams. The thesis also aims to generate actionable knowledge related to variants of online ASB that can be utilized by social, health and government organisations to handle crises associated with such behaviour.

1.2 Research Problems

The primary research question for this thesis is as follows:

How behaviour traits of an individual can be accurately and automatically identified from his/her use of different online social media platforms?

The amount of time we humans spend online in general, and on social media platforms in particular, has been growing in the past two decades. What started as an occasional online visit to check emails has expanded to most of the activities we perform in our daily lives. We have become dependent on online platforms for entertainment, work, business, socializing, shopping, organizing events, education, news and much more. Young individuals, who have grown up in this digital age, cannot imagine living without online platforms and tools. It can sometimes be hard for them to believe that most of these technologies did not exist some twenty years ago. And yet, we have become so reliant on these that missing them for a day can put a halt to most of our day-to-day activities such as socializing, entertainment and communication. Since most of our activities have moved online and especially onto social media platforms such as

Facebook, Twitter, LinkedIn, Instagram, emails and other online business tools, our interactions with the outer world have subsequently moved to these platforms. The outer world in this context constitutes our friends, family, colleagues, managers, subordinates, clients, suppliers, businesses and government officials.

We humans exhibit our behaviour patterns and personality traits when we interact with the participants of our society and social structures. Since most of our daily interactions with the people around us have moved online, the traditional ways to study behaviour and personality traits are not necessarily relevant and sufficient for today's digital world. Social scientists have primarily relied on surveys, questionnaires and interviews to collect data for behaviour studies. These methods are quite time consuming, cumbersome, expensive and in most cases less relative in the digital world in which we exhibit most of behaviour traits. Thus, the question for this research is to determine how behaviour traits of individuals in our society can be accurately and automatically identified and studied from the use of different online and social media platforms.

The above key research question is split into three related sub-questions and these are listed below:

Sub-Research Question 1: What sort of behaviour traits of individuals can be depicted from their online interactions on social media platforms?

In a broad sense, human behaviour can be divided into three categories: personality traits, states of mind and behaviour traits. Both behaviour traits and states of mind are dynamic in nature. This means that they may fluctuate quite often depending upon circumstances and situations. Personality traits on the other hand are relatively stable and are part of a human being for a longer duration of time. They do not change much with changes in circumstance and situation.

When we interact with other people and businesses online, we display a mixture of all of these three traits. We exhibit these by our written and spoken language, and by posting videos, images and text. Some behaviour and personality traits are easier to depict from online platforms than others. This difference may be due to body language, expressions and the tone with which we utter our words, and these may sometimes be difficult to depict in an online setting, However, the distinction is shrinking due to our increasing use of online video posting and conferencing. The aim of this sub-section of the research question is to discover the behaviour traits that can be depicted with high accuracy from our social media interactions.

Sub-Research Question 2: Can antisocial behaviour be identified from online discourse, and what machine learning, deep learning and natural language processing techniques and algorithms yield the highest accuracy?

Social media brings to our life its own benefits and advantages. It facilitates access to information, keeps us connected to the world, helps us discover new products and services, enables our work, and much more. Along the way, it can also incite unreasonable and unacceptable behaviour. One such behaviour is ASB which is a pattern of violation and disregard of the rights of others. It encompasses failure to conform to social norms, disrespect for others, lack of regret and remorse, reckless disregard for others' safety and aggressiveness and impulsivity. Online ASB prevents a lot of people from participating online; missing the advantages and benefits that social media platforms offer. Some users can be bullied, teased, shamed, and even have their lives threatened. Such behaviour inhibits true participation and leads to dire consequences such suicide. So, on the one hand social media brings benefits, and on the another, can become a breeding ground for ASB. Some social media platforms have measures that prevent nudity on their platform, and recently most have been trying to prevent the spread of disinformation in the form of fake news and fake media, however, no platform has seriously looked into curtailing ASB. Currently, if a user experiences such behaviour, he/she is expected to report it to the platform so it can take appropriate action. In most cases, victims are reluctant to report such behaviour due the fear of retaliation by the preparator and, in some instances, do not know how to report such behaviour. This leads to a lot of such behaviour going undetected; encouraging preparators to carry on. Considering the sheer number of global social media users and the data they create, manual detection does not make much sense and is not practical at scale.

This sub-section of the research question aims to discover techniques that can be used effectively at scale and with high accuracy, to detect ASB on online platforms. It also aims to investigate what natural language processing, machine learning, and deep learning techniques can yield the highest accuracy and precision.

Sub-Research Question 3: Can sub-categories of online antisocial behaviour be depicted with high accuracy, and what other knowledge related to online antisocial behaviour can be derived from the discourse and further utilized to curtail it?

Online ASB manifests itself in different ways. Some of the most common are lack of remorse, indifference, disregard for safety, consistent irresponsibility, aggressiveness and irritability, impulsivity, deceitfulness and unlawful acts. Some of these forms are more devastating for victims than others. ASB in its entirety is not at all acceptable, however the forms that are most treacherous should be addressed with utmost priority. The implication of not dealing with such threatening behaviour can send victims into hiding, depression and, in rare instances, push them to suicide. In the previous sub-section of the research question, the aim of the research was to depict ASB on online platforms with high accuracy and precision. This sub-section aims to go a step further, intending to not only depict ASB on online platforms, but categorize it into its most common forms. The question aims to establish that whether cutting-edge machine learning and deep learning techniques can be utilized to depict the different types of ASB manifestations at scale, bearing in mind the sheer number

of users of these platforms worldwide. This can help platforms deal with such behaviour based on its severity; dealing with the most disastrous first. The question also aims to discover knowledge related to such behaviour that can be further utilized to curtail it. Such knowledge can not only be useful for online social media platforms, but also for government and mental health organizations needing to come up with strategies and plans to tackle the online manifestations of ASB. This is imperative considering the amount of money and other resources that these organization spend in an effort to eradicate such behaviour from online platforms and our society.

1.3 Hypotheses

The primary research questions and its three related sub-sections are based on the following hypotheses for this study:

1. Social media behaviour is a proxy for a real-world behaviour. Since most of our daily activities, ranging work to entertainment and everything in between, have moved online, the behaviour exhibited while performing these activities by users/individuals can be taken as proxy for their real-world behaviour.
2. The social behaviour and personality traits of an individual can be depicted from his/her online activities. For Example: narcissism, paranoia, obsessive-compulsion, ASB, shopping habits, political views, livelihood preferences, etc.

3. Machine learning and deep learning algorithms using natural language processing techniques are able to identify online posts containing elements of ASB with a high accuracy from profoundly informal social media corpora.
4. Sentiment and semantic analyses of posts extracted from social media platforms can reveal useful and actionable knowledge related to various human behaviour-patterns and how the majority of online users feel about a particular event or a thing.
5. State-of-the-art deep learning algorithms can perform better than the majority of the traditional machine learning algorithms on natural language processing tasks when working with short text. The majority of online posts are considered short-text as these often consist of a few informally written sentences.

1.4 Thesis Contribution

Contributions of this research project and their significance have been detailed in the following paragraphs. The overall framework for the study is outlined in Figure 1.1.

Detection of social behaviour patterns from online platforms

Human behaviour and personality have always been topics of interest across research communities. It is imperative not only to social scientists, anthropologists, and psychologists, but also to government and healthcare organizations for their prevention and policy work. The psychological traits of

a person can be divided into three categories: personality traits, states of mind, and behaviour traits. Both behaviour traits and state of mind are dynamic in nature and personality traits are relatively stable. The three form the sides of an analogous behaviour triangle.

Traditional methods for behaviour and personality studies have primarily focused on qualitative research methods that constitute surveys, questionnaires and interviews. These techniques are quite cumbersome and time consuming when conducting large-scale behaviour studies. People around the world spend more time online than they ever did. Most of our interactions (work, banking, social and shopping) with the outer world take place online and this is where we humans exhibit our personality and behaviour traits. This makes online platforms and the data obtained from them, one of the best alternatives to, and supplements for, behaviour studies. To find new ways to explore human personality and behaviour patterns in the age of big data and AI is more imperative than ever. This research project contributes to the body of knowledge by proposing and implemented a novel approach to study social behaviour patterns. Information from online platforms was extracted and leveraged to detect and understand personality and behaviour traits.

Antisocial behaviour post detection from online corpora

ASB can often go unnoticed online due to the lack of appropriate systems needed to detect and eradicate it. Most of the manual procedures put in place by majority of the platforms fail to be effective at scale due to the manual

interventions required. Most of the time, preparators escape attention because victims either fail to report or platforms fail to take action because of their inability to detect such behaviours automatically. Detecting and eliminating online ASB is more crucial now than ever considering the large number of individuals, especially young individuals, who are advertently or inadvertently exposed to it. Online ASB is a social problem that demands urgent action. The problem contributes not only to monetary loss for governments around the world, but also to human loss due to a number of suicides related to it.

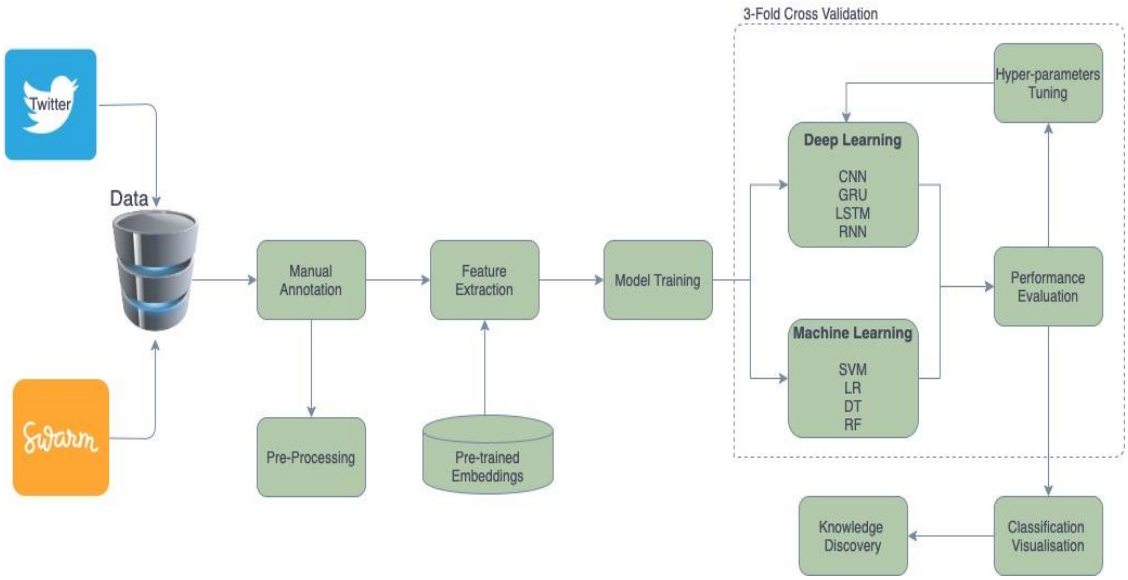

Figure 1.1 Overall Research Architecture

Gold standard benchmark datasets construction

There are datasets available related to domestic violence, cyberbully, natural disasters and mental health, however, a high-quality and an extensive annotated corpus for behaviour studies in general, and ASB in particular, is not yet available to the research community. Hence, novel gold standard ASB and Social

Behaviour datasets have been constructed for this research under the supervision of a domain specialist working actively in the area of ASB. Collecting data and constructing a dataset using manual annotation is an intensive, time-consuming and expensive process that requires significant of resources. Therefore, readily available and fully annotated corpora, related to both social behaviour and ASB, containing fine-grained information can lay the groundwork for effective future behaviour categorization studies.

Three different data sets: *Social behaviour, Binary classification, and Multiclass classification*, were constructed for this three-part research project. Each dataset constituted of collecting candidate posts, manually annotating them and discarding the posts that did not fit the criteria; all of which took an immense amount of time and effort. Availability of fully annotated public repository allows for rapid model training, validation and application implementation. These are also the first corpus for the specified tasks as there have been no such previously conducted studies requiring such datasets. Furthermore, these gold standard datasets that are available to the research community, and can be further enhanced and expanded for other studies and experiments with larger dataset needs.

Multi-class antisocial behaviour identification from social media

ASB appears online in many incarnations. Some are more detrimental than others. The automatic behaviour classification allows for efficient detection and

categorization of different types of online ASB. It affords an opportunity for platforms, health and government organizations around the world to take appropriate actions based on the category of online ASB detected. The approach is an extension to the preceding work on binary classification of online posts. Some types of online ASB are more of a nuisance than harmful, and automatic multi-class ASB identification can enable organizations to allocate resources where needed the most. Thus, multi-class behaviour categorization is presented, in which four distinct categories are identified to provide in-depth and fine-grained insights into the online ASB's severity and prevalence. Categorization of ASB can enable platforms and organizations to re-route their preventive efforts, improving the efficiency in detecting such behaviour online.

Descriptive statistical analysis of online posts

Descriptive statistical analyses were performed for social and ASB identification from online posts. For social behaviour, the analysis discovered various activities and related behaviours that individuals exhibit online. It shed light on how people behave when visiting different places, and at different times of a day. It also highlighted how behaviour is influenced by gender, weather and the day of the week. For ASB, in-depth statistical analyses were performed for both binary and multiclass categorization to facilitate knowledge discovery, and for comparison purposes. Particular words associated with a certain type of online ASB were discovered, along with the writing style of people who publish posts containing these words. Understanding of the average length of posts exhibiting

different ASB, types of words used in these posts, usage of stop words and punctuation, writing style and grammar all affords an opportunity to understand how preparators behave, the rationale behind their thinking, and manifestation of ASB online. All of these can assist in efficient feature extraction and better model training for future research initiatives in the area.

Semantic coherence analysis and knowledge discovery related to antisocial behaviour

Semantic coherence analyses using heat maps, word clouds, and Z-scores were performed in this research. Analysis investigated the relationship between the words that appeared in ASB posts and in non-ASB posts. It also delved deeper into analyzing words, that play an important role in assisting classifiers in categorizing posts into distinctive ASB categories, and the characteristics, relationship, coherence and semantics that these words share among themselves. Some words are more prominent in ASB tweets than others, and their occurrences in these posts are manyfold compared to the others. Some of these ASB words also appear in non-ASB posts, and most of these posts exhibited sarcasm. Highlighting the significance of these words in ASB and non-ASB posts was crucial to understanding how to efficiently detect and classify distinctive behaviours online. Some of the ASB words have higher correlations with other ASB words and some to non-ASB words. Whenever the latter combination appeared in a post, it pointed more towards other behaviours including sarcasm, rather than ASB. The insights discovered during these

analyses contribute significantly to the existing body of knowledge.

1.5 Thesis Outline

The remaining parts of the thesis are organized as follows:

- ***Chapter 2*** contains a literature review and outlines the background on behaviour studies conducted using social media data. It also provides background information on ASB, its aetiology, online manifestation, repercussions and obligation to restrain. The chapter looks at behaviour studies from a social and psychological science perspective and discusses how online platforms can inadvertently harbour and encourage unacceptable behaviours.
- ***Chapter 3*** presents a comprehensive overview of text mining and various operations within. It discusses the traditional machine learning and deep learning algorithms and how these are applied to a natural language processing workflow. Other computational techniques are implemented in this thesis are also presented in this chapter.
- ***Chapter 4*** proposes a methodology to conduct large scale human behaviour studies by utilizing data from social media corpora. Different people use social media for different reasons and while doing so exhibit behaviour traits which can be captured and analysed using various machine learning techniques. The chapter investigates behaviour patterns based on weather, gender, day and the time of the day. It explores various categories of

activities in which people indulge, depending on different independent variables. The chapter includes the construction of a medium-scale data set from Twitter and the Swarm app. Sentiment and psychometric analyses are conducted to gauge how people feel and react when moving around and visiting disparate places. Statistical analyses, topic exactions, text clustering and language analyses were performed to establish behaviour patterns and for knowledge discovery. Statistical analyses included the study of spatio-temporal and geo-temporal user activity patterns from the dataset. The chapter demonstrates the effectiveness of the methodology for large scale behaviour studies by producing meaningful results.

- ***Chapter 5*** presents a data-driven approach to classify online posts that contain elements of ASB. First, a benchmark dataset of candidate posts is constructed with labels for antisocial posts and non-antisocial posts. Posts were annotated manually under the supervision of a clinical psychologist who specializes in ASB. Textual features were then extracted from predominantly unstructured natural language textual data for further processing. State-of-the-art deep learning algorithms were experimented with for the text classification process, and to establish the model with the highest performance on most evaluation metrics. The performance of the developed approach was evaluated against the performance of the traditional machine learning algorithms and various feature extraction techniques on the underlying dataset. Furthermore, semantic coherence

analysis was performed utilizing word clouds, Z-scores and heat maps to generate knowledge related to the prevalence of ASB online.

- ***Chapter 6*** (Multi-class/subclass detection and classification of antisocial behaviour) introduces a deep learning-based framework for multi-class ASB post categorization. The automatic content classification approach addresses the issue of scalability that is imperative when dealing with the amount of online social media data. The chapter provides fine-grained insights into distinctive categories of online ASB that are prevalent on most platforms. First, four separate categories of ASB are identified based on the DSM-5 [1] guidelines. Second, candidate posts are collected and the benchmark corpora is constructed. Third, state-of-the-art deep learning algorithms are trained on this dataset. Accuracy, precision, recall and f-scores are compared for the four different deep learning architectures used in this chapter. These are also evaluated against the evaluation metrics obtained from the four different traditional machine learning algorithms by experimenting on the same dataset with two different feature vectors. Furthermore, visually enhanced interpretation in the form of scatter plots and confusion matrices, is provided for a better understanding of the classification process. Fourth, error analysis Is performed to investigate inaccuracies in the classification process with the potential of reannotating posts, and retraining the model to further improve performance accuracies.

- ***Chapter 7*** concludes the thesis by summarizing the findings from each chapter. It highlights some of the limitations of the study and provides recommendations for improvement and future research directions.

CHAPTER 2

BACKGROUND

This chapter outlines the background on social media analytics, behaviour science research on social media and antisocial behaviour. Furthermore, it discusses the current state of research on online ASB, its aetiology, online manifestations, and repercussions. ASB is discussed in the context of social media platforms. Applications of social media data in behaviour science provides an overview of what has and can be achieved using such data. Finally, examples of research studies utilizing user-generated online content are presented as they offer actionable knowledge for time- and health-critical issues.

2.1 Social Media Network Data Analytics

Social media has transformed information consumers into information producers. This phenomenon has enticed researchers (from various disciplines) to study online social media as an important source of data to explore human behaviour in the physical world [6, 7]. Social media generates unprecedented amount of data and has led to new ways to discover urban functions and human behaviour.

By aggregating millions of check-ins on platforms such as Swarm and Facebook Places, researchers can reveal distinct visit patterns and busy times for various locations within a city [8]. Based on this information a recommendation system

can be developed to provide real time information on interesting events and their statistical deviation from past trends [9]. Aggregated check-ins can also reveal positive and negative sentiments about the places people visit. This can be achieved by conducting real time sentiment analysis of the data, and information can then be fed back to a recommender system to make recommendations [10, 11]. Based on these recommendations, people can plan their outings and visits.

Discovering similar trends from data collected by traditional methods, also called the 'top-down approach', for real time functionality is not possible. Data obtained for urban planning by top-down approaches such as remote sensing, nationwide surveys, and geographic information system has its limitations. It is intricate and fails to reveal the complex dynamics of a city. It is also difficult to extract the emotions, perceptions and experiences of people using this approach [12]. Even though such data has been used extensively in the past to study urban functions, urban planning and geographical information systems, it comes with limitations related to the time and effort required for collection, processing and analysis.

On the other hand, data generated by 'bottom up approaches' offer superior alternatives and include user created data in the form of online blogs, check-ins, and social media posts [13]. Researchers are not required to create questionnaires or a surveys; normally sources of very rich data and information

[11]. Instead, data can easily be collected using techniques such as APIs, web crawling, software tools such as 'Ncapture', and web scrapping. By analysing comments, posts, images, etc. posted by users of social media, researchers can infer important information about them. This may include the type of art people buy, the theatre they visit and fashion they favour [14].

User generated social media data is not only useful in discovering human behaviour, is also useful in generating city knowledge [15, 16]. Vernacular geography encapsulates the spatial knowledge that people use to visualize and communicate the places around them on daily basis. Through the data posted on social media directly or through metadata, people unknowingly share a body of knowledge about their surrounding geographic world. Flickr, which people use to post their photos, provides meta data that helps to understand vernacular geography of an area [17, 18]. It is not uncommon for people to add tags to the images they store on the platform and, in most common scenarios, they use the location at which the image was taken as a tag. Vernacular geography also concentrates on how people name and demarcate places in everyday use, and the knowledge gathered from this can be utilized to develop a geographical information system [19]. There is no limit to the type of questions one can answer using the same set of social media data and use cases for the data depends on researchers' imaginations.

User activity modelling has been an area of interest for researchers in the field of pervasive computing. For developing an activity model, researches usually rely on data from different streams such as location sensors and smartphones with inbuilt GPS sensors [20, 21]. The data gathered is then analyzed to develop a 'user behaviour model' to discover content or location dependent knowledge. This knowledge is then fed into a recommender system for targeted advertising and marketing. In particular, such information can be analyzed to determine user activity duration.

The activity duration can then lead to other types of predictive analytics. Based on the past time spent at a location, a system can suggest further places of interest to a user. A location based social network, such as Swarm, offers a new and different avenue for data to develop a user activity model [22]. This data can be sufficient on its own, without the need to collect data from other sources to model human activities. Not all check–ins qualify to be used for such system however, with millions of check–ins every day, they can be filtered to select the ones that meet the criteria, and the number can still be substantial. A lot of people check-in regularly when they travel from one place to another. For example, they check–in once they leave the office, and again when they reach a restaurant or other place of interest. They may check-in again when they reach home after a visit to a place of interest. To develop such a system, only consecutive check-ins from same users can be analyzed. The methodology would be to select all such tweets in an area of interest for which a user activity

model has to be developed. Once collected these check-ins can be analyzed to estimate times spent at different locations. While doing this, it has to be kept in mind that the time required for travelling between two locations must be considered. A user can walk, drive or take public transport to reach a next destination. So, the time spent between destinations has to be estimated and deducted from the total time spent at venues. One easy approach would be to take an average of time taken by walk, train and car. Another approach would be to find the distance between two locations, and if the distance is shorter than a pre-defined distance, chances are that the user walked to the next location (for example, it does not make sense for a user to catch a taxi or to drive to a location, which is just 200-300 meters away). A software program using Google's location API can also be written to estimate the time between locations. Once the travel time is estimated, the time spent at a venue can be calculated. Based on the time spent at a particular location, a user's interest and liking for that place can be deduced to develop an efficient activity model [23].

2.2 Social Media and Applications of Classification Techniques

The interest in data obtained from social media platforms and studies based on it has been steadily growing over the last few years [24]. This is mainly due to the increasing number of social media platforms, their types, and the growing number of users on these platforms. Interest from the business sector and government agencies has also contributed significantly to the growth and usage

of social media platforms. Numerous types of studies are being conducted on social media data and some examples of these are opinion mining, sentiment analysis, topic modelling and emotion modelling. Primitive applications of opinion mining and sentiment analysis primarily focus on blogs and certain web wages [25].

Applications have expanded into other areas such as movie reviews and recommender systems which use different machine learning techniques. One such study by Pang et al. [26] experimented with the support vector machine and the naive bayes classify movie review into negative and positive and carried out a performance evaluation of the two algorithms used. Numerous similar studies have been conducted for sentiment analyses of new product releases, government decisions, and music releases using natural language processing, machine learning and deep learning, and the applications of such studies are growing by the day [27]. Amolik et al. [28] experimented with classifiers. Naïve bayes and Support Vector Machine, and feature vectors to conduct sentiment classification of reviews of Hollywood and Bollywood movies with a very high accuracy. Pak et al. [29] constructed a tweet corpus from Twitter to conduct linguistic analyses for online posts using Naïve Bayes, and classified posts into neutral, negative and positive categories. Other similar studies using traditional machine learning algorithms, namely Naive Bayes and Support Vector Machine and natural language processing techniques to conduct sentiment analyses were able to achieve high accuracies [30, 31]. Studies that have implemented deep

learning architectures to conduct sentiment analyses on data collected from Twitter, were able to utilize automatic feature extraction, and context incorporation in order to achieve superior performances [32, 33].

Applications of classification from social media data are also making their way into healthcare. A study by Nivedha et al. [3] successfully attempted to classify tweets from Twitter into healthcare and non-healthcare related categories. Two different classifiers, namely the Decision Tree and the Naive Bayes were experimented with, with Decision Tree outperforming the latter. Collier et al [34] implemented classification techniques on social media data to Identify syndromic categories using related keywords from health ontology. Support Vector Machine and Naïve Bayes were implemented for performance evaluation. Paul et al. [35] used ailment topic aspect modelling to categorize tweets that mentioned anything related to a disease, its symptoms and its treatment. Unigram, bigram and trigram features were used to implement a Support Vector Machine with its linear kernel. A study [36] successfully detected flu-related tweets using a Support Vector machine and unigrams. In another study, which touches a different domain, Ramya et all [37] used a Support Vector Machine and C4.5 classification algorithms to identify aspects of any advocacy from social media data. Traditional machine learning algorithms were implemented on children's and women's healthcare related data for the research project.

In the same domain of healthcare, a study identified mental health disorder-related tweets using state-of-the-art deep learning architectures [38]. The classification was performed to detect the signs of anxiety, bipolar disorder, and depression from users' social media activities. A multi-class classification approach was carried out. A number of studies have been conducted using deep leaning and machine learning techniques to detect and classify different crisis situations from online platforms, and propose response solutions to these crises [39-42]. A study by Imran et al [43] used automatic text classification techniques to achieve the same. A similar study by Nguyen et al. [44] used Convolutional Neural Networks on social media data to detect and classify crisis situations online, and propose response solutions.

Aryo et al [45] identified the issue of spam in tweets, and carried out a study to successfully classify spam from non-spam posts. In behavioural economics and finance, Bollen et al. [46] used deep learning, sentiment analyses and text classification techniques to gauge public sentiment towards certain equity in stock markets, and how sentiments towards these equities impacted their prices in the succeeding days. To conclude, this section demonstrates the current research and applications of text classification techniques used on social media data in various domains, laying the ground for future work in this area.

2.3 Social Media Data for Social and Behaviour Science Research

Personal traits and behaviour patterns of individuals can be discovered from their web search history, web browsing history and bank statements [47, 48] however, due to strict privacy restrictions, this information is not available to all. For example, apart from the bank with which a person has an account, only some government organizations can have access to a person's account statements. A bank statement can tell the type of places a person has visited and the type of products he/she has bought at physical and online stores. A lot of behaviour patterns can be learned from such data and can be used to offer new products, personalized search engines and targeted marketing services. However, due to privacy regulations, even the banks that own this data cannot use it for the abovementioned purposes.

A person's web search history and browsing history is only available to some (the search engine company and to some developers) and hence cannot be obtained freely by everyone for social research. However, publicly available data set such as check-ins (from Swarm and Facebook Places) is easily available to all and can be used for the same. Not every user of Foursquare and Facebook Places share his/her data, but the vast majority do. Some users, who do not like to share their activities with others for research purposes, keep their settings private. This prevents their data and activities from being seen by others and to be used for research by the platform and other researchers. Only a small number of users keep their profile private because the whole idea for these users to be

on the platform is to share information about the things they do and the places they visit. This provides a vast source of data for behaviour and social research.

People like to be associated with the organizations, places and activities that they enjoy, and this is exactly what platforms such as Swarm and Facebook Places enables them to do [49-51]. By checking into a particular nightclub, a user broadcasts his preference for that place. Similarly, by checking into a particular fashion brand store, a user does the same for that brand. Other traits such as ethnicity, sexual orientation, religious beliefs and food preferences can be discovered from similar check-ins. A study by Kosinski et al. [52] considered exactly this. The authors explored how the private traits and attributes of a person can be predicted from his/her digital footprints. In their study, the researchers used 58000 volunteers to discover that Facebook likes can be used to accurately and automatically predict a wider range of highly sensitive personal attributes including religion, sexual orientation, political views, ethnicity, happiness, use of addictive substances, intelligence, gender, age and parental separation. The model that the researchers developed correctly differentiated between heterosexual and homosexual men in 88% of the cases, Republicans and Democrats in 85% of the cases and Caucasian Americans and African Americans in 95% of the cases. The researchers concluded that a model trained with appropriate social media data can accurately predict many behaviour traits from a wide variety of personal attributes. This conclusion indicates the need for more

research in the area to discover other aspects of human behaviour by using a dataset that is publicly available to all.

2.4 Social Media Data for Predictive Analytics

Social media has become a credible source of data for predictive analytics in both the business and academic worlds [53-55]. Asur et al. [56] offer an example of how this data can be used for such predictive analytics. They extracted 2.89 million tweets (140-character messages on Twitter) referring to 24 different movies released over the period of three months. They built a simple model showing that the rate at which Twitter users talk about a movie can correlate with its box office success. Researchers were able to predict box office revenue for these movies based on the conversations on the Twitter platform. Not only did they predict these revenues with significant accuracy, their predictions outperformed market-based predictors.

Chipotle, one of the major food chains in the US, lost 6% of its share value after a 30% fall in first quarter sales of 2016 [57, 58]. Some investors were caught off guard, but not all. Foursquare (the parent company of Swarm) had, a few weeks earlier, predicted the same drop in sales. Apparently, there was an E.coli outbreak at some of its locations that led to a fall in visits made to its restaurants. Foursquare saw a significant drop in check–ins at Chipotle's restaurants throughout the US. Based on the analysis on the number of check-ins, it announced (on 12th April 2016) that Chipotle's sales would drop by 30%. The

accuracy of the prediction by Foursquare, in regard to a percentage drop in sales, surprised everyone.

This was not the first time Foursquare was successful in predicting sales with such accuracy. In 2015 it predicted Apple iPhone's sales, leading to the launch event of its new 5, 5s and 6 iPhone models. Based on the check-ins (implying the foot traffic at Apple stores across the US), Foursquare predicted that Apple would sell 13 – 16 million phones during its launch weekend. Apple ended up selling exactly 13 million phones during that weekend [58]. This prediction was a further evidence of the growing use of check-in data and other social media data for predictive analytics. The way things are going, with more and more people joining such platforms and producing data at an unprecedented rate, social media data won't just be an alternative source of data for predictive analytics, it will be the main source of such data.

2.5 Other Use Cases and Applications of Location Based Data

Researchers in social media have explored location-based data for some other applications. One such study by Cheng et al. [59] used a large data set of 22 million Foursquare check-ins from 220,000 different users. Researchers initially collected a relatively small number of check-ins from these users and then dived deep to extract up to 2000 check-ins from each of these users. They looked at mobility patterns and human behaviour from three different aspects: repeated check-ins at the same place, whether the social status of users defined the places

they visited (check-in) and the sentiment of those users (extracted from messages associated with check-ins) when interacting with the places they checked into. It was discovered that most people check-in at places with a higher frequency initially and the frequency declines as time passes. For example, someone catches a train to work every morning from a local train station and check-ins every day. The temporal relation shows that as the time passes the user's check-in frequency at the train stations goes down. One explanation of this could be that initially the user is excited to check-in at a same place every day but, with time, it becomes monotonous.

With regards to social status, researchers studied areas with high net income residents and compared them to areas with low-income residents. They found that people in high-income areas take more distance trips when it comes to travelling. The explanation for this could be that affluent people have more resources to travel. Another noteworthy observation was that people living in dense areas such as New York City had more trips but the trip sizes were small when compared to trips by people who lived in the countryside and less busy cities. The author thinks that people living in the countryside do not usually have access to a lot of places to go for entertainment and hence they do not go out much. However, when they do go out, they have to travel far which is the opposite of what people living in dense neighbourhoods such as New York, Los Angeles and San Francisco do.

Conducting sentiment analysis on the comments associated with check-ins, researchers found that most of the comments were neutral in nature with very few associated with positive or negative emotions. The results indicated that most people do not express strong emotions to venues they check into. Researchers found patterns of check-in during the day and during the week. 9 am, 12 pm and 6 pm are the busiest time when it comes to people checking in during weekdays, however the pattern is a bit different during the weekend when the evening is the busiest time. The number of check-ins increases as the week passes from Monday to Friday, with Friday evening being the busiest. Explanation for this could be that most people finish their working week and feel relaxed and go out a lot and check-in at venues. The study also compared check-in patterns of three different cities: New York, Los Angles and Amsterdam. It showed that people in Amsterdam are early starters when it comes to check-ins. This might indicate that people in Amsterdam start work earlier than most people in the two US cities.

Cranshaw et al. [60] explored mobility patterns in Pittsburgh and validated their results using a qualitative approach in which they interviewed 27 Pittsburgh residents to see how their perceptions of the city projected onto research findings there. The results for both the qualitative and quantitative study were very similar; again validating the significance of using Location-Based Social Network (LBSN) data for research. All of the above discussed research projects have used some sort of social media data to analyze and to explore various

human behaviour related activities and traits. A lot of these studies have been conducted to explore mobility patterns in various cities around the world. [61-63] However, none of these studies has considered weather as a moderating variable that can impact behaviour patterns. Weather can influence how people interact with, and express emotions to, places they visit [64]. The aim of this study is to explore and extract human behaviour and personality traits from LBSN and other online platforms. To best of my knowledge, this research project is the first to explore such behaviour patterns from social media platforms.

2.6 Behaviour Issues Online

An attempt to understand and regulate content and user behaviour online must commence with the understanding that some online communication is partly regulated by laws, public policy and the platform's own policy. The present online ecosystem has led to the rise of asymmetric and opaque relationships between platforms and their users. As a result, it is reasonable to ask whether platforms will endeavor to accept responsibility to nurture an environment that promotes and values users' wellbeing. Online media, more than any other media, encourage and permit active human behaviour such as interactions and searches for information [65]. The behaviour that users exhibits are, to the largest degree, contingent on social and environmental cues and structures [66]. Seemingly small aspects of digital ecosystems can nature individual behaviour and scale it up to a noticeable collective behaviour change. Mostly, major social

media platforms have taken the role of intermediary between publisher of content and content readers. More than half of the world's Internet users turn to social media platforms for their information needs; making it a prominent part of their lives.

These platforms have the potential to offer digital cues that assist users in assessing the epistemic characteristic of the material posted, and how it can impact them individually [67, 68]. However, platforms currently have difficulties differentiating between normal text and extremist content because both types of messages are often tagged with similar keywords. Another general shortcoming of these platforms is their epistemic and endogenous cue quality assessment that require an understanding of the many issues related to human behaviour. A single approach may not protect every vulnerable user online as every online platform is different in the way it disseminates information and enables interaction. Some are predominantly video based, some image based, and others text based. Unacceptable and extremist behaviours can take any of these forms, therefore demanding tailored policies and measures [69]. Another type of platform that experiences challenges regulating user behaviour are gaming platforms [70]. These platforms enable game playing at global level, affording a different medium of interaction for users: some aggressive and antisocial, and others vulnerable. Even though most of the research shows that similar numbers of women and men play games on these platforms [71], women are generally perceived as more susceptible to unacceptable behaviour and are

often harassed and targeted with toxic behaviour [72]. Gaming platforms have been the most inequitable of the platforms; exposing women to toxic and harmful behaviour [73].

Although men are more likely to come across online harassment, such as being embarrassed, being insulted and made fun off, women tend to experience surprisingly severe forms of harassment such as threats, stalking and, in some instances, sexual harassment [74]. Other research has demonstrated that gay, lesbian, transgender, bisexual, non-white and minority groups are also targeted at higher rates [75, 76]. The ever-growing popularity of, and reliance on, these online platforms, makes toxic online culture and cyber aggression an increasingly serious social problem that needs a well thought and effective preventative action plan. Cyber aggression has been shown to be strongly related to suicidal ideation of victims, compared to traditional bullying, and many suicides have been linked to victims being exposed to cyber aggression [77, 78].

2.7 Antisocial Behaviour

According to the 'Diagnostic and statistical manual of mental disorder' (DSM-5), a diagnostic tool used by mental health professionals [1], there are ten known personality disorders. These disorders are categorized into three clusters based on their similarities and prognosis. ASB is one of ten personality disorders and falls in a cluster alongside Borderline personality disorder, Histrionic

personality disorder and Narcissistic personality disorder. It is a lasting pattern of behaviour that diverges significantly from the expectations of society. It is usually inflexible, pervasive, and leads to impairment and distress. A person displaying ASB violates and disregards the rights of others without considering any implications.

2.7.1 Aetiology of Antisocial Behaviour

Understanding the aetiology of ASB may be the first step towards preventing and eliminating it. ASB disorder is a part of the Cluster B of personality disorders including borderline, histrionic and narcissistic personality disorders. Individuals who suffer from these behaviour personality disorders appear emotional, dramatic and erratic. These characteristics are common in all four disorders in the cluster. A person with ASB often displays disregard for other people's emotions and feelings, and often engages in activities that are considered illegal however, the manifestation of such activities dwindles as the person grows older [1].

There may be many elements leading to the development of ASB. Some of these are genetic influences, maternal depression, parental rejection, physical neglect, poor nutrition intake, adverse socioeconomic situation, and socio-cultural factors. [1, 79-83]. These factors can be categorized broadly into three main categories: neural, genetic and environmental [84]. Antisocial behaviour due to neural factors has been studied through structural and functional approaches. Structural studies assess the brain's morphology, and functional studies assess

its activities. Together these studies try to understand core neural regions that are related to salience detection, affect and controlled cognition, including the frontal cortex, amygdala and anterior cingulate cortex [84]. The genes of a person linked to ASB may develop during adolescent [79]. Certain types of gene combinations are closely associated with such behaviour. A child who is raised by biological parents diagnosed with ASB has a high probability of developing ASB. Some studies, however, have concluded that if the same child is raised by adopted parents who do not suffer from such a disorder, he/she has a lower chance of developing ASB [85, 86]. Therefore, genes play an important role in the onset of ASB, however the impact can be mitigated if the individual's environment can be changed into a more positive one.

Some environmental factors that may trigger or lead to antisocial behaviour are exposure to community violence, family dysfunction and peer influence [84]. Research has shown that being part of a disadvantaged community, living in a poor neighbourhood, being dependent on social security, being a part of female-headed household or not having a job, may exaggerate or trigger the onset of ASB [87, 88]. Being a part of a broken family, facing maltreatment by either parent, or a parent's mental health can also impact an individual's mental health. Apart from parents, maltreatment by others, can also prompt him the manifestation of ASB [89]. Child neglect, in general, has been associated with ASB [90]. The sort of company an individual keeps usually influence his/her behaviour and personality, and vice-versa. So keeping company with an

individual who manifests ASB can lead people to display such behaviour as well [91-96]. It is known that smoking is harmful, not only for the person who smokes, but also for individuals around him who inadvertently and passively inhale the smoke exhaled by a smoker. Many studies have linked maternal smoking during pregnancy and severe mental disorders, in particular ASB, in offspring [80, 97-104]. Similar to smoking, excessive parental drinking is also associated with an offspring developing ASB.

As well as neural, genetic and environmental factors, some studies have found a link between poor quality nutrition and childhood ASB [83]. Deficiency of B-Vitamin is particularly is linked with ASB and other mental health and behaviour disorders [105]. So, there are many factors that can lead to an individual developing mental health disorders in general, and ASB in particular. Here, we have discussed some of the crucial factors leading to ASB and since this remains an active area of research, we may learn more about this personality disorder in the future.

2.7.2 Manifestation of Antisocial Behaviour

Antisocial behaviour emerges in various forms online. Some of the most common are trolling, cyberbullying, threatening, hostile behaviour, offensive language and the publication of inappropriate images. Trolling is widespread on social media, and magazine and news websites. Trolls are general visitors to a website who write offensive and inflammatory comments in the public section.

Their main aim is to disrupt an online discussion and grab some attention in the process. They disregard the author of the writing on which they comment and also show no respect to other commenters. They do this by posting comments that are sexist, hateful, racist and profane in nature. Trolling intensity ranges from subtly provoking someone to outright threatens and abuse [106]. For some, trolling traits are inborn and they have a history of trolling and engaging into such behaviour online. These people seem to experience enjoyment at the cost of others [107-109]. This type of trolling is associated with sadism [110]. For others, environmental variables, situation, and context can come into play [111]. A negative mood and seeing other people trolling online can also prompt people to troll [112]. A person who otherwise has a very pleasant and normal personality can sometimes be pushed into getting involved in trolling inadvertently. This sort of situation may arise if someone, who is not a troll, feels that he has been pushed around and feels that he need to stand up to such behaviour. In the process the victim himself can start trolling the abuser in an attempt to prevent future trolling or to teach the troll a lesson [112].

Studies have found that trolls focus their effort on a small number of threads and make issues of petty things. They usually write worse things than people who do not troll and, in some instances, their posts are irrelevant to the topic of discussion. Males are more likely to get involved in trolling compared to females [113]. Overtime these trolls become less tolerated by the online community and are reported and removed from the conversation, and in some cases from the

community altogether [114]. The impact of trolling on victims can sometimes be more devastating than if they experience similar behaviour in real life [115]. Exposure to online trolling can lead victims to experience psychopathological outcomes such as anxiety, depression and low self-esteem [116].

While trolls mainly focus on being a nuisance and attracting attention, cyberbullies target individuals. Instead of posting generally offensive and inflammatory statements in the public comment section of a website, they post abusive and vicious comments about a single individual. Cyberbullying refers to the use of an online platform such as Twitter, Reddit, Facebook, to intentionally and repeatedly harass or harm an individual [117]. Cyberbullies focus on intimidating, shaming and demeaning their victims. Unlike trolls, the cyberbully does not usually want to attract attention and instead focuses more on targeting an individual and causing them distress. To do this, they post images, text, audio and video targeted at individuals on a repetitive basis [117]. This media is abusive and aggressive in nature, and is intentionally drafted to bully someone online. As the use of online platforms has increased, so the cyberbullying and it is quite prevalent in school-age children. Depending on the measuring tool applied, 10%-40% of school-age children experience some sort of cyberbullying [118]. Cyberbullying has been receiving a lot of attention from government authorities and social scientists in recent years because of its association with a large number of suicides [119]. Trolling and cyberbullying are the two most prominent manifestations of online ASB. Threats, misleading and

wrong information, offensive language, sexism, racism, and the use of rude and taboo words, are some of the other ways ASB can manifest online.

2.7.3 Online Antisocial behaviour

Online ASB is a social problem and a public health threat. It is one of the ten personality disorders and entails a permeating pattern of violation of the rights of others, and disregard for their safety. It exists online in the form of aggression, irritability, lack of remorse, impulsivity and unlawful behaviour. The exhibition of online ASB appears to be a manifestation of everyday sadism. Online platforms can inadvertently encourage the proliferation of such behaviour by affording culprits access to other online users. Without having any measures in place to restrain such behaviour, online platforms leave a vulnerable group of people at risk. ASB can prevent this vulnerable group of people from lawfully going about their lives, and prevent them from exercising their right to social participation.

Exposure to online ASB affects a lot of individuals and inhibits their genuine participation on social media platforms. Ramifications of such behaviour can propel a vulnerable individual to take extreme actions, and in some instances commit suicide. Apart from sadism, boredom, desire to cause damage, revenge and attention-seeking are some of the other motivations that have been linked to such behaviour [109]

To discourage such online behaviour, the current measures taken by social media platforms are often not enough to curtail it. These measures often require victims to manually notify platforms of such behaviour [120]. The approach is not scalable and often fails as most victims are reluctant to report due to the fear of retaliation from the preparator, and hence most incidents go unnoticed. These online platforms encourage freedom of speech but fail to draw a line between ASB and freedom of speech. Platforms can also sometimes inadvertently encourage the proliferation of such behaviour by affording culprits access to other online users. Without having measures in place to restrain ASB, online platforms leave a vulnerable group of people at risk.

2.8 Chapter Summary

This chapter provides a brief overview and background on recent work related to social media and text classification, using traditional machine leaning algorithms and state-of-the art deep learning architectures. The research studies discussed in this chapter are from different domains including, healthcare, social science, finance and computer science. The studies afford an opportunity to understand the current state of research related to online platforms and related applications. The chapter has also provided a number of examples, from a variety of domains, of raw social media data extraction and transformation into valuable knowledge using machine learning and deep learning. Most of the research and techniques discussed have been utilized to implement real-world application. The chapter also provides seminal information related to the

current state of behaviour and personality studies on social media data. The findings from the overview are significant and form the foundation of this thesis.

CHAPTER 3

COMPUTATIONAL TECHNIQUES

The preceding chapter provided a background to social media analytics and applications, and the prevalence of ASB on these platforms. The current chapter presents a comprehensive overview of the computational techniques used in this thesis.

Social media platforms provide an abundance of user-generated data in the form of text, images and videos that can be explored for knowledge discovery. Knowledge derived from these platforms, and applications based on this knowledge, can be utilized in areas such as healthcare, social science, political science and criminal justice. Despite the apparent advantages of data exploration from social media platforms, there are a number of challenges associated with collecting and processing such data. The current chapter presents an overview of techniques that can be leveraged to manipulate social media data to extract useful patterns and knowledge. The chapter also discusses the various methods within these techniques that can be fine-tuned to achieve optimum results.

3.1 Text Mining

Text mining is a technique for extracting meaningful associations and knowledge from a collection of unstructured datasets. Knowledge discovery and data mining are terms synonymous with text mining. A typical text mining

workflow consists of collecting the right type of data, pre-processing of the data, applying various analysis techniques, recognizing patterns and associations, evaluation and interpretation of the discovered knowledge, and visualization.

The first step in text mining, also known as information retrieval, involves identifying a suitable and appropriate dataset. The data can be collected automatically using an API (application programming interface) and crawling technique from a number of sources in a certain format or can be manually collected using a questionnaire or form. Depending on the type and amount of data required, an appropriate method can be used to construct the data set. Before the data can be analyzed, a number of pre-processing steps are taken to clean and transform the data into a form that can be easily understood by tools and algorithms. Some of the important pre-processing steps are discussed below.

3.1.1 Segmentation

This is a process often used to identify sentences and paragraphs in a large piece of text or transforming text it into meaningful units such as topics, words and sentences.

3.1.2 Tokenization

Tokenization is one of the key aspects of any natural language processing task and is commonly used with count vectorizers and deep leaning architectures. It is a technique which separates text into small units known as tokens. The tokens can be characters, words or sub-words.

3.1.3 Normalization

Normalization is a process of transforming a chunk of text into a standard form. For example, converting all letters in a piece of text into the same case (upper or lower). This is a crucial process, making it easy for algorithms to learn from text, as both upper case and lower case letters have distinctive codes in coding systems used by machines or algorithms to transform and understand natural language.

3.1.4 Stop Word Removal

Some words do not contribute much to a text's meaning, and such words and can be removed to simplify the learning process for machine learning algorithms. Words such as 'a', 'and', 'of' and 'the' may not end up adding a great deal of value, context or meaning and can sometimes be removed before the machine learning training process is undertaken. Depending upon the nature of an application, stop words are sometimes left in the text if their value is deemed vital in the training process.

3.1.5 Stemming and Lemmatization

Both stemming and lemmatization are applied to generate a root from of a word. Stemming truncate suffixes and tenses of a word and shrinks it to its stem, whereas lemmatization recognizes and reduces a word to its base. Unlike

stemming, lemmatization uses the WordNet corpus to produce lemma leading to a relatively inferior performance to stemming.

3.1.6 Pruning

Pruning assists a machine learning model to be faster and smaller by removing weight connections and increasing inference speed leading to a decrease in model storage size. Pruning encourages the removal of superfluous parameters from a model that are over parameterized. There are terms in any document that appear either too frequently or too infrequently, often making an insignificant contribution to the performance of a model, and hence are pruned to reduce model storge size.

3.1.7 Treating Synonyms

This is a process of identifying two similar terms or words that have the same meaning and replacing one with the other without impacting the meaning and semantic of it. It is done to make a model lean and to improve its performance.

Once the pre-processing steps are completed, a piece of text is transformed into a representation that the model can understand. Therefore, the text is encoded using a language model such as word embedding, bag of words or n-gram.

3.2 Feature Extraction

Feature extraction assists a machine learning model in selecting a set of relevant features that capture the character of the text. The technique saves the space and time complexity of machine learning algorithms and enhance their outcomes by avoiding overfitting and reducing variance. The extraction process can be sub-divided into syntactical analysis, morphological analysis or semantic analysis. Some examples of applications that use feature extraction are sentiment classification [121], opinion analyzer [122] and automatic online post classification [123]. These three sub-categories are briefed described below.

3.2.1 Syntactical Analysis

Syntactic analysis explores and present a logical meaning for a piece of text. In this technique, the correct rules of grammar and exact meaning are considered in defining the logical meaning. It investigates text for meaningfulness by comparing it with grammar rules. Parsing and part-of-speech (POS) tagging techniques constitute syntactical analysis [124]. Parsing determines a syntactic structure of a piece of text and analyzes its constituent words in line with the grammar rules of an underlying language. The outcome of the process is a parse tree in which a sentence is represented as the root, and noun and verb phrases as immediate nodes. The words in the sentence forms the leaves of the tree. Conventional parsing approaches are based on probabilistic, statistical and machine leaning techniques. Tools such as OpenNLP and Stanford Parser are

often used for parsing. POS tagging, on the other hand, is defined as a process in which one POS is assigned to a word based on its definition and context. The techniques commonly used for POS tagging are probabilistic methods, rule-based methods, lexical-based methods, and deep learning based methods.

3.2.2 Morphological Analysis

Morphological analysis deals with stemming, stop-word removal, and tokenization [125]. Stemming is the process of removing suffixes and different forms of a word and reducing it to its root. For example, the words ‘flown’, ‘flies’, ‘flying’ can be stemmed to their root ‘fly’. Some examples of stemming algorithms are affix-removal, suffix-stripping, successor variety, brute-force and n-grams [125]. To reduce data sparsity and to standardize the terms in a sentence, Porter stemming is often applied. Stop words often do not add significant meaning to a text and can, therefore, be eliminated. Stop words in the English language constitute of pronouns, prepositions and articles, and some examples are ‘the', 'an', 'at' and 'a’. In most cases, removing stop words leads to improvement in the classification process by shrinking feature space and decreasing data sparsity (206).

Tokenization aids in interpreting a piece of text by dividing it into small chunks, known as tokens. When the tokenization is applied to a document or a paragraph to get sentences, it is known as sentence tokenization, and when a [126] sentence is spliced into single words, the process is called word

tokenization. Keras, Gensim and NLTk are some of the libraries and methods that can be utilized to perform tokenization. In addition to the above, morphological analysis also includes eliminating links, non-textual symbols, non-ascii characters, hashtags, numbers, user mentions, and punctuation.

3.2.3 Semantic Analysis

In Linguistics, semantic analysis is the process of relating syntactic structurers to the writing as a whole from all the levels of sentences, phrases, paragraphs and clauses [127]. It involves a wide variety of processing techniques that extract facts, concepts, events and attributes from text. The two most widely used approaches for semantic analysis are based on machine learning and rules. The machine learning approach utilizes statistical analysis and the statistical co-occurrence of terms by developing relationships between words within a document. Whereas the rule-based approached works with entity extraction and is one of the oldest approaches used for NLP, the rule-based approach tends to focus on pattern-matching and needs support from dictionaries. SentiWordNet [128] and WordNet-Affect [129] are the two most widely used approaches for the this purpose. WordNet-Affect is a linguistic resource and is commonly applied to derive lexical representation from affective knowledge, whereas SentiWordNet is a lexical recourse for mining public opinion and is open source.

3.3 Feature Selection

Feature selection is another important step in text mining, and follows feature extraction. This step includes forming a vector space to improve accuracy and scalability. The primitive function of any feature selection task is to pick up the most important and decisive features [130]. For efficient classification to take place, selecting the right subset of features in the feature vectors is imperative, bearing in mind the high dimensional feature space used for most text analyses tasks. Selecting the appropriate and good features preserve the semantic and original meaning of a text, improving the classification accuracy.

A number of studies have explored the role of the right kind of features and their ability to aid in the classification process. One such study [131] explored how to choose a superior and compact subset of features to reduce the cost, utilizing maximal statistical dependency. Another approach [132], known as the fast feature technique (which relies on conditional mutual information) has also been widely implemented. The approach promotes mutual information utilization among selected features to ensure that these are both minimally depended and individually informative. The approach by Michalcea et. al. [133] investigated numerous procedures to define semantic similarities among collection of words and sentences in a text. This approach is based on simple lexical methods such as latent semantic analysis and pointwise mutual information.

Two other widely used feature extraction techniques are TF (term frequency), also known as word frequency, and TF-IDF (term frequency inverse document frequency). TF works by counting the number of times a term appears in a document to determine its topic information. TF-IDF works slightly different from TF, calculating TF and IDF separately and then diving TF by IDF (TF/IDF). TF in TF-IDF is calculated as mentioned, whereas IDF measures the weight of an underlying word in a document. The more often a word appears in a text, the smaller the weight allocated to it, and vice versa. The idea behind this technique is that if a word appears frequently in a text, it may contribute less to the overall meaning and topic of a text [134]. The TF-IDF has successfully demonstrated superior results when applied to sentimental analysis tasks [135] and to recognize name entity in micro posts [136]. These techniques have also been successfully implemented in other classification tasks such as to classify noun phrases from Twitter [137] and for POS tagging to overcome noisy and sparse data, on Twitter [138].

3.4 Text Mining Operations

This is the algorithm implementation and kernel stage during which a specific application of text mining is undertaken to discover patterns and knowledge. The algorithms implemented can be related to machine learning, data mining and artificial intelligence. The choice of algorithm is made based on the

underlying expected outcome such as classification, clustering, topic detection, etc. These text mining tasks serve the needs of commercial, political, medical and academic research. The widely used applications of text mining in academic research and implemented extensively in this thesis are described below.

3.4.1 Clustering

Clustering algorithms automatically ascertain different types of texts/documents into distinctive groups and categories, also known as clusters. The documents and the texts in a particular cluster resemble significantly, in character and context, the others in the same cluster, and vary or contrast with the documents in other clusters (categories). Clustering algorithms investigate and identify disparate features within a text, and these form a basis to classify the text into a specific cluster. Clustering algorithms mainly fall under unsupervised algorithms and do not require training to perform. They do not require prior knowledge of text and are widely used in searching documents, customer feedback analysis, recommendation systems, customer segmentation, etc.

3.4.2 Classification

Classification algorithms mostly fall under supervised learning and require a significant amount of training before they can be used in a model. A classification algorithm requires prior knowledge about the document it is applied to. The knowledge is acquired during the training process in which the

algorithm learns, from the feature, how to assign a text to a class or category. The training phase consists of two steps: training and testing. A typical dataset is sub-divided into training and testing subsets. A technique such as k-fold is often utilized for the task. The algorithms are trained on the training dataset and tested on the testing dataset. Performance is measured using evaluation metrics such as accuracy, F-score, precision and recall [139]. The majority of the text analysis tasks conducted using data obtained from social media use classification algorithms.

3.4.3 Topic Detection

This is a technique in which latent topics within a document are revealed using algorithms. The technique is very useful when working with large text documents by reducing the need for manually reading and identifying topics of interest within a document or set of documents. Similar techniques can be utilized to summarize a document and automate the process. Apart from its use in academic research, the technique can be utilized in auxiliary diagnosis, medical systems, clinical decision making, news and online social media to detect the important underlying themes, topics and summaries. Topic extraction has become one of the vital knowledge discovery approaches and is inextricably bound up with classification and clustering techniques.

3.4.4 Sentiment Analysis

As the name suggests, this technique involves detecting the sentiment, whether positive or negative, of a document. The technique can investigate and gauge

human emotions in the form of positive and negative feelings. The feelings are generally related to the underlying topic of a text. Several tools and techniques can be used to enhance the application of sentiment analyses to further subcategorize positive or negative sentiments to their various degrees. Sentiment analyses can be applied to text at a sentence level, paragraph level, document level, and at a large corpus level. [140]. It is a very active area of research in both academia and industry, with industry leading the charge with the number and types of application. Some examples of application in industry are customer reaction to a new product release, sentiment for the state of a particular stock in stock market, sentiment detection about a topic from social media, sentiment detection from feedback, sentiment detection from a new movie release. The analyses obtained from sentiment analyses can be further utilized for prediction applications.

3.4.5 Psychometric Analysis

Whereas sentiment analysis depicts positive and negative sentiments in their variant degrees from a document, psychometric analysis deals with other form of behaviour traits. Depending on a research question or a business application, a model can be trained to depict many human emotions and behaviours. Some examples of such emotions are anxiety, eagerness, fear, threat and anger. Psychometrics analysis is a relatively new area of research and has amassed much interest after its successful application and research in the area of sentiment analysis. If the research question is seeking just positive and negative

emotions, sentiment analysis is applied, however, if the need for emotion seeking is complex and cannot be explained by just positive or negative emotions, psychometric analyses are often implemented.

3.5 Visualization

Once text analyses are completed using any of the aforementioned techniques and approaches, there is often a need to interpret results and findings in a form that is palatable and easily understood. This is where visualization comes into play. Most of the patterns, relationships and theories are better understood using visualization. It assists in driving meaningful findings and conclusions from a set of complex results after implementing text analysis techniques. Visualization assists in straight declarative and simple presentation of results, aiding effective decision making. Depending on the research area and the business application, different forms for visualization techniques are utilized ranging from simple bar, graphs and pie charts to word maps and gauge charts. Visualization is a very active area of research and has application, not just in text analysis, but in many other research area. Techniques such as metaphors, data storytelling, data journalism, mobile friendly visualization, visualization through AR (Augmented Reality), VR Virtual reality, real-time visualization, etc. are making their way into visualization and are being actively researched. Figure 3.1 shows the visualization of prominent words that users of social media write to exhibit ASB online.

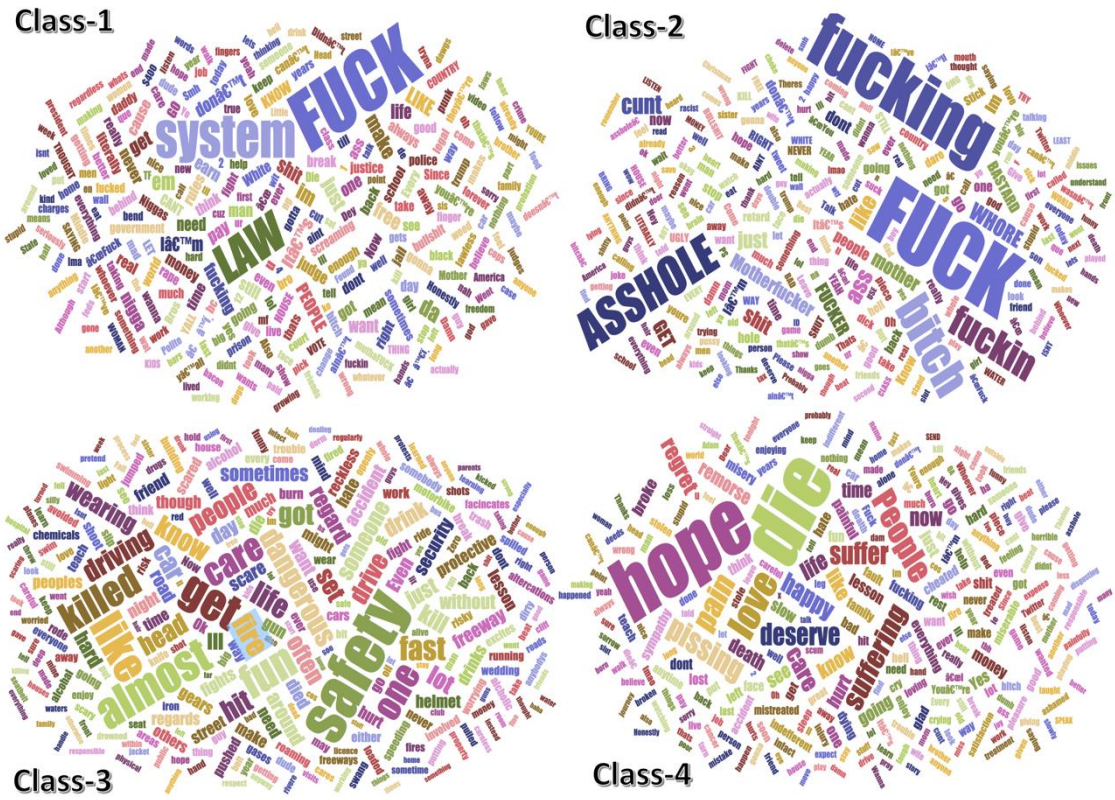

Figure 3.1 Word cloud for antisocial behaviour classes visualization

This figure presents four different classes of ASB with each cluster representing one class. The word cloud helps in understanding the context of each class without digging deeper into the dataset from which these words are taken. Figure 3.2 is a graph that presents training cycles of four deep learning algorithms using two different word embeddings, namely GloVe and Word2Vec. Visualization makes it easier to understand how each algorithm performed during each training epoch. Trying to obtain the same information from a table filled with numbers may not always be as effective.

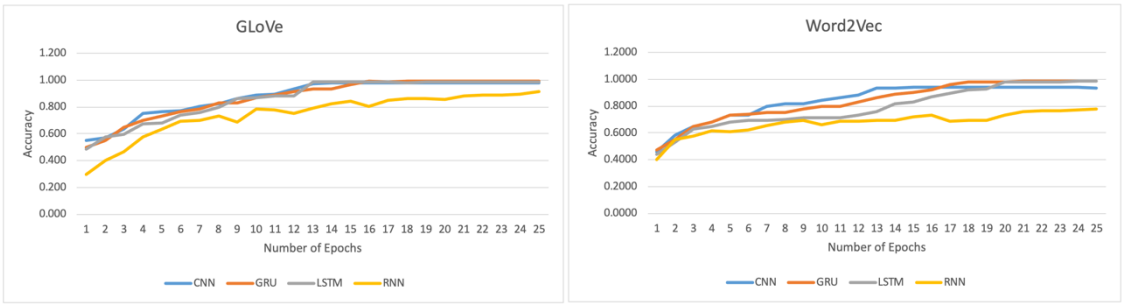

Figure 3.2 Epoch graph for word embedding performance visualization

Similarly, scatter plots in Figure 3.3 helps in understanding the classification processes of RNN, LSTM and GRU; the three different architectures of the deep learning technique. The algorithms were given the task of classifying the dataset into five different classes of ASB with each color representing a distinct class. It can be seen that the clusters in GRU are more defined representing better classification performance. Other visualization techniques that have been implemented throughout this thesis that help in understanding the results and concepts better.

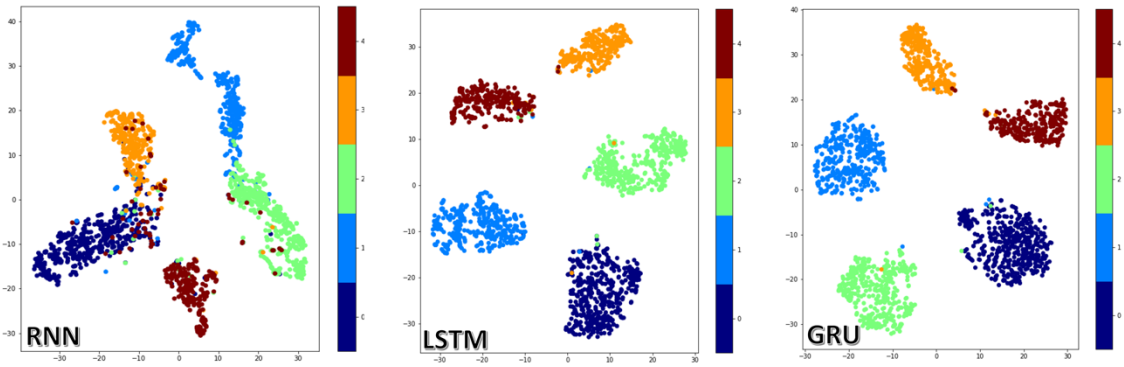

Figure 3.3 Scatter plot for classification visualization

3.6 Text Mining and Related Terminologies

Text mining is a broader concept and an umbrella term often used interchangeably with information retrieval, natural language processing and data mining. Below is a brief description of some of the most commonly used terms that lie at the intersection of text mining.

3.6.1 Natural Language Processing (NLP)

NLP techniques are related to the interactions between natural human language and computers. The natural language here does not just imply 'English' language. It implies any of many different human languages spoken by humans in verbal speech or text format. The language could be German, French, Mandarin, Spanish, etc. Research in the area initially started with rule-based techniques in the 1950s and has now moved more towards machine learning and deep learning due to the higher accuracy and performance with the later techniques. Applications of are techniques are plenty and some examples are inter-language translation, text classification, text summarization and intent classification.

This research area has primarily been the focus of academics in computer science, and linguistics however, with advances in machine learning and deep learning technologies, it has attracted additional interest from industry with new consumer technology products such as echo devices, google now devices'

etc. being introduced on regular basis. NLP has made its way into many domains including healthcare, in which medical practitioners have integrated speech to text tools in their workflows.

3.6.2 Information Retrieval

Information retrieval lies at the intersection of NLP, machine learning and text mining. The technique discovers and presents appropriate information relative to a user's requirement within a certain context. It enables users to search for a vast amount of data (text, image, audio and video) to obtain information of interest. In other words, it is the science of searching for required information from a document or set of documents. One of the most important and universal application of information retrieval is the web search, where a user finds information of interest after having submitted a query to a web search engine. The returned information does not have to be in text format and can also be in images, audio and video formats.

3.6.3 Statistical Analysis

Statistical analysis is a process of collecting and interpreting data to uncover trends and patterns. The technique is utilized in statistical modelling, research interpretation, designing research studies, etc. Statistical analysis falls under quantitative methodology or conducting research and deriving information. Numerous techniques fall under statistical analysis and these can be grouped under 'descriptive statistics' and inferential statistics. Descriptive statistic

summarizes data from given samples utilizing indexes and techniques such mean, median, mode and standard deviation. Inferential statistics, on the other hand, draw conclusions from a dataset taking into consideration the random variations. It draws conclusions using methods such as regression analysis, confidence interval and hypothesis testing. Prominent methods to accomplish the aforementioned are z-score, t-value, chi square, etc.

3.6.4 Data Mining

Data mining is a process of discovering correlations, associations, patterns, anomalies and outlier detection in a large dataset using a number of techniques. It is an interdisciplinary subfield of statistics and computer science and has gained traction with the availability of large datasets, relatively cheap storage, fast processing power and improved algorithms. The patterns, associations and correlations, utilizing data mining, are often discovered in order to predict future trends and outcomes, and to discover new knowledge for research and decision making. Data mining is a term often used synonymously with text mining, which mainly deals with text data (linguistic), however, data mining encompasses numerical, image, audio, video, signal and any other types of data. The techniques used in text mining are often borrowed from data mining. The area is one of active research in academia and industry alike. Some of the applications of data mining in business are price optimization, prediction, customer segmentation and fraud detection.

3.6.5 Machine Learning

Machine learning relates to the study of computer algorithms. These algorithms largely utilize statistics to discover relationships and patterns in a large data set. The term data encompasses all types of data including numerical, linguistic, images, videos, signals, geographical and spatial. Since machine learning deals with computer algorithms, all of this data must be in digital form. Machine learning and the related algorithms can be further sub-divided into three main categories: supervised, unsupervised and reinforcement learning. Supervised learning consists of two parts: training and testing. In the training part, algorithms are trained to perform a particular task. The training data set is labelled, and algorithms learn from these labels utilizing features within the data set. Once trained, the algorithms are tested on a testing dataset to gauge their accuracy, using performance evaluation metrics such as accuracy, F1, precision and recall. However, in unsupervised learning, no training is required and algorithms are applied to data to discover the patterns In a dataset. In this form of machine learning, no labels and annotations are required as algorithms attempt to discovering previously unknown patterns. Therefore, the technique is considered to be self-organized learning.

In contrast to supervised and unsupervised learning, reinforcement learning is a technique in which algorithms react to their environment on their own and learn from it. The point to note here is that, unlike supervised learning, the reinforcement learning algorithms learn without any labelled data. The

algorithms work on two different states. When the learning starts, they are on start state and need to determine how to reach the end state. During the learning process, algorithms do get some sort of feedback on how they are learning from their environment. The technology has many real-world applications in systems supporting driverless cars, self-navigating vacuum cleaning, self-navigating lawn mowers, etc.

3.6.6 Artificial Intelligence

Artificial intelligence (AI) began in the mid-20th Century and was aimed at building efficient cognitive systems to assist a variety of human activities and the automation of workflows. This underlying requirement has become the ultimate aim for all the recent research and development. The initial techniques were rule-based, followed by fuzzy logic. Recent progress in the areas has mainly been due to the incorporation of machine learning and deep learning techniques in AI research and implementation. It is fair to conclude that most of the recent advancements in AI research has been based on deep learning technologies. It is during its recent history (last 10-20 years), that AI research has accelerated, leading to various industrial application. The applications are in diverse fields such as robotics, healthcare, banking and finance.

3.7 Overview of Machine Learning Algorithms

The following is an overview of the machine leaning algorithms used in this research.

3.7.1 Logistic Regression

Logistic regression in its rudimentary form is a logistic function that has been utilized to model a dichotomous variable. The algorithm has been around since the start of the 20th Century and was initially used in biological science research and later made its way into social science and computer science. The algorithm is appropriate when the dependent variable is binary/dichotomous. Logistic regression, similar to the other regression analyses, is utilized primarily for predictive analyses. It has many complex extensions that can be used to model a classification problem into a multi-class categorization however, its main use has been in binary classification where a dependent variable has two separate values such as cat or dog, pass or fail, up or down, win or lose; and these are denoted by indicator variables with the binary values labelled as '0' and '1' [141].

For multi-category classification, multinomial logistic regression is often utilized. If these multi-categories are in order, this kind of logistic regression is known as ordinal logistic regression. Therefore, the statistical model is utilized to describe a relationship between a single dependent dichotomous variable and single or many ordinal, nominal and ratio level independent variables. The log-odds, in the model, for values labelled '1' are the linear combination of one or many predictors (independent variables). The independent variable in logistic regression can also be continuous, not just a real number of values. Logit is the unit of measure for log-odds that is derived from the logistic unit, representing

an alternative name. One thing to be careful of when considering logistic regression for use in a classification task, is the actual model fit. Adding more than necessary independent variables to this statistical model will likely lead to an upsurge of variance related to the log odds that is expressed as R^2 inadvertently resulting in model overfitting. An over fitting model often works well on the dataset it is trained on, however it fails to generalize.

3.7.2 Support Vector Machine

The support vector machine (SVM) is one of the most reliable and commonly used machine learning algorithms in which vectors are data points that are closest to the hyperplane (decision surface). These data points are often challenging to classify and impact directly on the optimum position of a decision surface. While learning on a training dataset, SVMs come up with a number of different candidate final solutions and, from those, find the optimal solution. The primitive aim of the algorithms is to maximize the margin that lies around the candidate separating hyperplane. The separation margin (d) are the margins between the hyperplane and its closes data point for a given bias b and the weight vector we. The optimal hyperplane with the maximum margin is the one for which the separation margin d is largest.

The decision function in SVMs is specified by a subset of training samples called support vectors. This leads to the problem becoming quadratic, making it relatively easier to solve using standard methods. The separation data point in

two dimensions is basically a line (in its simple form), however transforms into a hyperplane for higher dimensions [142]. The simple idea behind the efficient working of a SVM is that it picks the optimum among a number of possible solution hyperplanes by maximising the margin between the training data points. Support vectors are the critical elements of a training set that can lead to a change in position of a dividing hyperplane when removed. Therefore, support vectors play a critical role in the training set. The issue of finding an optimal hyperplane can be considered as an optimization problem that can be resolved using optimization techniques.

3.7.3 Random Forest

Random forest (RF) is a supervised learning architecture and can be used for both classification and regression problems; making it a broad scope algorithm. The algorithms often use the Gini index (a formula that decides node on branches of RF) for classification problems, and employ the mean squared error when used with regression problems. The algorithm is composed of disparate decision trees, each carrying the same nodes but distinct data leading to different leaves. The decisions from multiple sub-tress are merged to find the final answer to the underlying problem. The final outcome of the algorithms also represents the averages of all the outcomes from the sub-trees. In other words, the algorithm grows multiple classification trees.

When an input vector is put through the algorithms, it passes that vector on to the multiple sub-trees for their classification outcome. Once the outcome from these sub-trees is obtained, their votes for the classification are aggregated and the outcome which received the highest votes is the final outcome [143]. The trees in the RF grows as follows: a) If N is the case number of the training set, these are sampled at random; replacing the original data, b) If the number of input variable is M and the number m < M. The number M is specified in such a way that each node, at random, m variables are chosen leading to the best split of m in use node splitting. As the forest grows, the value of 'm' is kept constant, c) The trees are not pruned and are grown to their largest possible extent.

The error rate of the forest is depended on the correlation between trees and the strength of individual trees. The error rate increases with increased correlations. Furthermore, the forest error rate decreases with an increase in the strength of an individual tree leading to a tree, which has low error rate, being a strong classifier [144]. By reducing the size of m, both strength and correlation can be reduced and vice-versa. The optimal range of m is often found in between. The out-of-bag (OOB) error rate is often used to find the appropriate value of m and the range is decided. It is the adjustable parameters of a RF. RFs are rarely over fitted and relatively fast to train and test. However, at the same time they can be relatively slow to make predictions once trained, and the algorithm must be aware of the missing values in dataset and outliers. The algorithm is efficient on larger datasets and can easily handle a large number of input variables without

deletion. They can also calculate proximities among pairs of cases, often used in locating outliers, clustering, and provide fascinating aspects of a dataset.

3.7.4 Decision Tree

A decision tree is a supervised traditional machine learning algorithm used for predictive analysis. The top node is called a root node and branches that originate from the root signify distinctive options. Leaves in a decision tree represent labels (classifications), non-leaf nodes represent features and the branches of the tree are aggregations of features that help in classification process. [145]. A typical decision tree is a recursive partition of feature space into subspaces that form the bases of a prediction. Internal nodes in a decision tree are the ones that have outward edges. The remaining are the called terminal nodes. Decision trees carry on classification utilizing a fixed set of ordered decisions on a feature set. Decisions carried on by the internal nodes of a decision tree form a split criterion. A class is assigned to each leaf in a decision tree and the small disparities within a training set can lead to distinctive splits and hence distinctive decision tree. Therefore, a resulting error can lead to a large variance for a decision tree.

Due to their non-linearity, decision trees are considered relatively flexible in exploring and predicting many candidate outcomes. Disregarding some of these can then take place. A smaller tree is often considered efficient to build and run. A common heuristic used for this purpose is ID3 and is grounded on gathering

information. ID3 is proceeded by its enhanced version called C4.5. To prevent overfitting in decision trees, overfitting pruning is often utilized enabling a tree to generalize better. While building a decision tree, deciding which attribute to have at the root can be a complicated task, and selecting any random variable to do the job does not lead to the optimal outcome. Criteria such as information gain, entropy, Gini index, reduction in variance, gain ration and chi-square can often be used address the issue. The value for each and every attribute can be calculated using these criteria. Following this, the values are organised, and the attributes are positioned on a tree in order of their values, i.e. the attributes with the highest value will be at the top/root. The attributes are assumed to be continuous with the Gini Index, however categorial with information gain. Entropy represents randomness in the information and the higher entropy leads to exertion in extracting conclusions from the information. [146, 147].

3.7.5 Naive Bayes

Naive Bayes algorithm is based on Bayes theorem and is a probabilistic machine learning algorithm. The algorithm works by assuming that each input variable depends on all the other variables. Conditional probabilities of each and every variable is then changed to different conditional probabilities, leading to class labels allocated to variables. The next steps consist of multiplying all these independent conditional variables. The multiplication is performed for every class label, and the label which comes out with the highest probability is often

taken as final classification output. The rule is also referred to as the 'maximum posteriori' decision rule (MAP).

Bayes theorem, in its simple form, is widely implemented for various classification modelling tasks and is commonly known as 'Naive Bayes'. The algorithm has proven quite effective and, hence, is generally used for document/text classification problems [148, 149]. The words within a piece of text are encoded as binary, frequency or count vectors, Gaussian, binary or multinomial distribution. The three commonly used variants of Naïve Bayes are: 1) Gaussian Naïve Bayes that utilizes Gaussian distribution, 2) Binomial Naïve Bayes that is based on binomial distribution and 3) Multinomial Naïve Bayes, as the name suggests, based on a multinomial distribution. The variant is often implemented based on data type. For binary classification, Binomial Naïve Bayes is often suitable, whereas for categorical variables, such as labels, counts, etc, multinomial is more appropriate. For numerical variables, such as measurements, Gaussian is often implanted. Furthermore, a dataset that has mixture of data types as input variables, may have to use separate types of distributions for each and every variable. It can be noted that using the aforementioned distributions, which are the most commonly used variants, is not necessary and there are other options available that can be utilized.

3.8 Overview of Deep Learning Algorithms

Deep learning in itself can be considered a sub-field of machine learning. Artificial neural networks have long been a part of the machine learning algorithmic approach and are inspired by the neurons in human brain and their biology. Unlike the neurons in the human brain, which can connect to any other neuron in human brain, artificial neurons form discreate layers. Neurons in one layer are connected to neurons in subsequent layers enabling data propagation. The idea behind the development of deep learning algorithms was inspired by the traditional artificial neural networks [150].

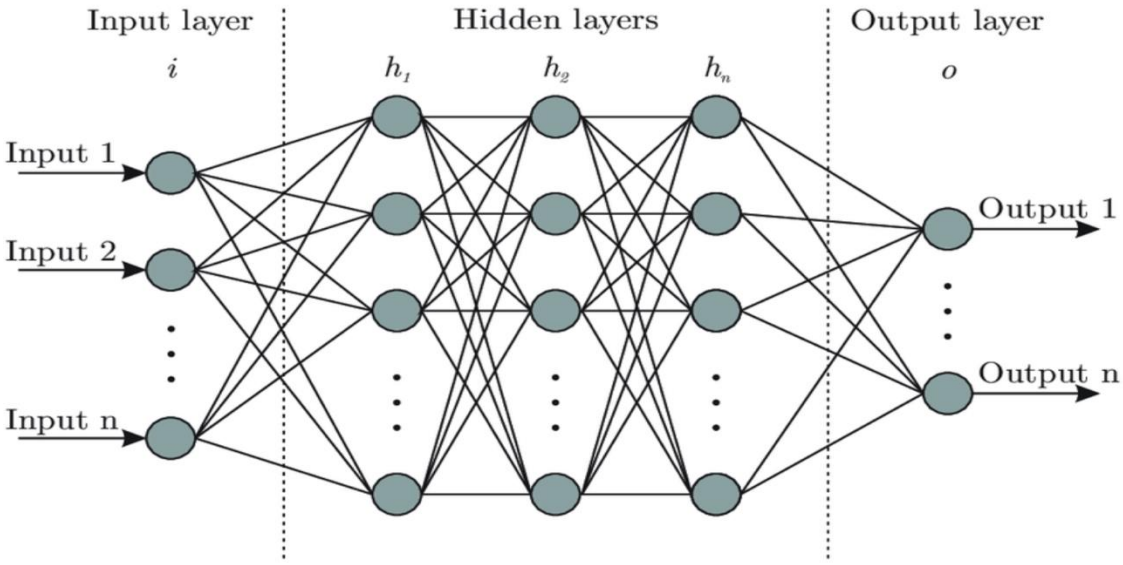

Figure 3.4 Deep learning architecture

In the context of traditional machine learning, artificial neural networks have very few layers containing a few neurons, however in deep learning, the number of layers is significantly higher and each of these layers consist of a larger

number of neurons. The rationale behind this is, by increasing the number of layers and the neurons within those layers, a larger amount of data can be processed for training and testing, thus enhancing learning capabilities. The architecture of a deep learning network is presented in Figure 3.4.

To begin with a deep learning algorithm, data is fed into the first layer of an architecture which can be broken down into neurons in that layer. The data then propagates to the next layers, and neurons within layers assign weights to their inputs. Weights are relative to a task and whether it Is performed correctly or incorrectly. These weights are then aggregated to determine a final output. An important advantage of using deep learning over traditional machine learning, is its ability to extract features automatically from a raw data set. Using traditional machine learning algorithms often requires having a specific domain knowledge which may contribution to feature extraction and usage. However, with deep learning's automatic feature extraction ability, this need is eliminated, enabling its practitioners to work in wider areas and tackle disparate problems[151].

A number of factors have propelled the growth in deep learning's development and usage and some of these are: the availability of large datasets, faster processing power of computers and the declining cost of data storage. Bearing in mind that the technology is still in its nascent form and has been around for only a few years, its applications have increased in almost every domain. Deep

learning is the key factor in the technology used in the recent progression in AI, and its applications in consumer and business technology. Driver-less cars, self-flying drones, automatic vacuum cleaners and lawn mowers are just some of the examples of AI applications. In finance, it has been used for anomaly detection, fraud detection and managing risk. In healthcare, it has been assisting radiologists to be more effective in examining larger numbers of images (x-rays, scans), enabling the detection of disorders and illnesses with signal processing, and making medical staff more productive by affording speech-to-text applications and tools. Implementation in behaviour studies is limited at this stage. In fact, this research project is one of the first to have used deep learning in behaviour studies and is the first to use it for detecting personality traits and behaviour patterns from social media platforms. Below is a brief overview of the deep learning architectures used in this research project.

3.8.1 Convolutional Neural Network

A convolutional neural network (CNN) is a type of deep neural network used predominantly for visual imagery. Its application in other domains, such as natural language processing, recommender systems, time series and brain-computer interfaces is not uncommon. Its implementation with natural language processing, due to its exceptional ability in sequent data analyses, is gaining momentum. The primary two operations of the network are convolution and pooling. The role of the convolution operation is primarily to extract features and feature maps from an underlying raw dataset, preserving its spatial

information. The pooling operation, which is also known as subsampling, then leads to the reduction of feature map dimensionality from the previous convolution operation. Max and average pooling are the usual pooling operations carried on in convolution neural networks. An example of a CNN architecture is presented in Figure 3.5.

CNNs are also known as space invariant or shift invariant neural networks, based on their translation invariance and shared weight architecture [152]. CNNs got their name from 'convolution', a mathematic operation that is employed in this network. This operation is used in at least one of its layers instead of the generally used matrix multiplications.

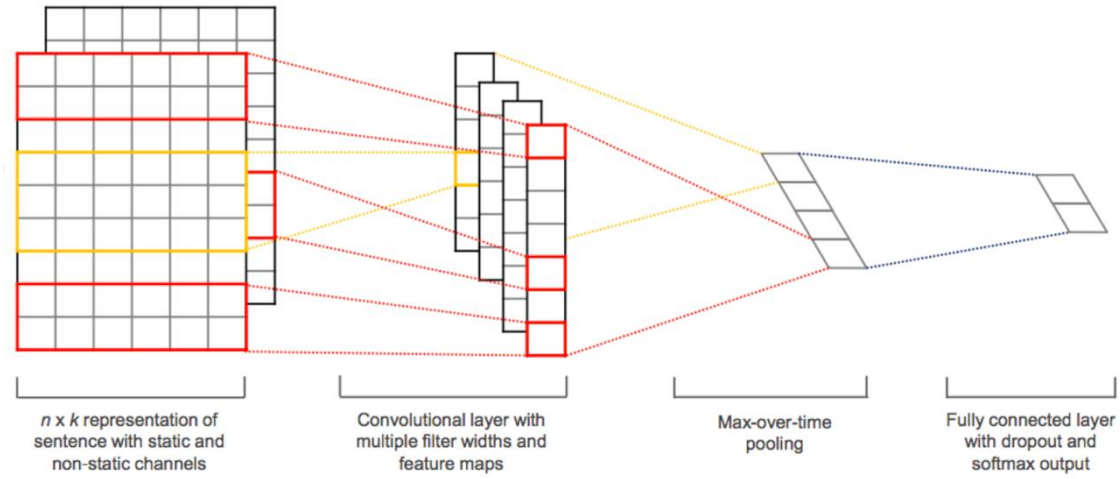

Figure 3.5 Convolutional architecture and inner workings

CNNs generally consist of the following four layers: input, convolutions, pooling and fully connected layers. The input layer is the first layer and is represented by a tensor/matrix and forms the building block of a CNN. It

accepts the initial data and related high-level features into the architecture for further processing, and plays a passive role as it is not fed with any feedback from a preceding layer. In general, subsequent layers are fed with weights and biases, however this is not the case with the input layer as it is the very first layer in a network.

The next layer in the architecture is the convolution layer. It uses filters to perform operations known as convolutions, and during the process scan the input 'I', related to its dimension. The hyperparameters of the architecture includes the stride 'S' and the filter size 'F' and the subsequent output 'O' is called an activation or feature map. The layer convolves an input and moves its result to the next layer in the network. Each and every neuron in CNN processes information related to its receptive field. For the filters within the convolution layer, it is imperative to understand meanings behind the hyperparameters used.

The pooling lawyer follows the convolution layer in a CNN, and is typically a down sampling operation with certain spatial invariance. Average and max pooling are distinct kinds of pooling in which the average and the maximum value is respectively taken. The fully connected layer, which connects each and every neuron in a particular layer to neurons of another layer, typically works on a flattened input. Fully connected layers in CNN consist of an activation function, and are generally used in optimizing certain objectives of a network,

such as class scores, and often present near the end of a network [152]. This research experimented with CNN due to its recent progression and high performance with natural language processing tasks, and because no similar study has implemented CNN for detecting personality traits from an online corpora.

3.8.2 Recurrent Neural Network

Recurrent neural networks (RNN) are a type of ANN that enables the preceding outputs to be utilized as inputs and have hidden states at the same time. RNN is one of the most powerful architectures that has been designed to work with sequences of data; making it a perfect fit for time series, stock markets, genome, text, spoken words and handwriting among others. The primary distinction between RNNs and other neural networks is that RNNs sequence and time components into account and possess a temporal dimension. RNNs are applicable to images and decomposes them into series of patches and treats them as sequence.

Analogous to the human brain, RNNs have two distinctive sources of information, the recent past and the present [153]. RNNs combine these two inputs to determine how they are going to proceed, forming their functioning. Unlike feedforward networks, RNNs ingest output from a feedback loop which is connected to preceding decisions as inputs, moment after moment. The mechanism makes RNNs work similar to the human brain utilizing its memory

function. This addition of memory to RNN has a crucial function and is utilized to capture information that lies in sequences, which is in contract to feedforward networks. This sequential information is captured in the hidden stage of a network and cascades through it and is utilized by the succeeding layers in processing new data points. In doing so, the network determines correlations between data points, which can be separated by several moments. The architecture of a RNN is presented in Figure 3.6.

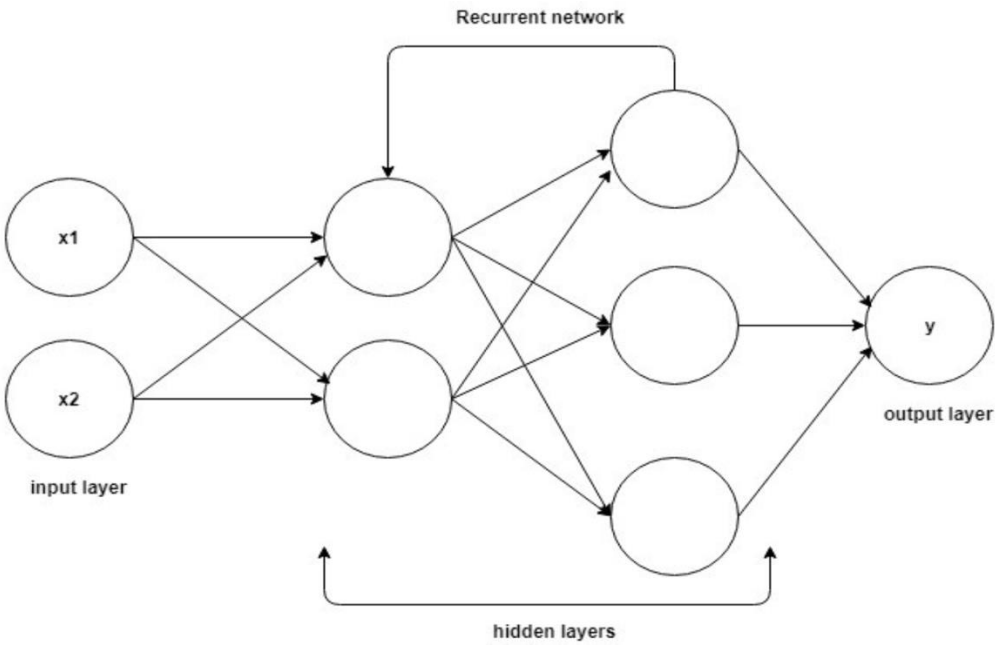

Figure 3.6 Recurrent neural network with a feedback loop

These correlations are known as long-term dependencies mainly because of the fact that the downstreaming of an event is a function of, and is dependent on, one or more of the several events preceding it. Another way of looking at RNNs is that they enable the sharing of weights over time.

Another crucial component of RNNs is the weight matrix which acts like a filter and determines the importance of both past and present information. The

backpropagation technique is utilized here to return the error generated and to adjust weights of each information set, and to reduce it to its lowest possible level. The aggregation of the hidden state and the weight input is compressed using either a tanh or a logistic sigmoid, making gradients feasible for backpropagation. As long as the inbuilt memory persists, every hidden state holds traces from both the last hidden and all preceding states. This becomes possible as the feedback loop appears at all the time stamps. The thing to note here is that the sole purpose of recurrent nets is to precisely categorize sequential inputs; the gradient descent and the backpropagation of error both facilitate this process. One crucial advantage of using RNNs over the other neural networks when working with sequential data, is that these can process inputs of any stretch and size without impacting the model size, and the weights are distributed over time. However, this can result in slower computation, and the task of retrieving information from preceding states that have long since passed, can become challenging.

When working with natural language processing tasks in general, and text in particular, keeping track of proceeding information is necessary to make full sense of the data. Even though RNNs can capture such information, vanishing gradients can sometimes become a limiting factor for superior performance, making parameter tuning and learning challenging [154]. To address this shortcoming, the gate mechanism was proposed, leading to the development of

Gated Recurrent Units (GRU) [155] and Long Short-Term Memory Networks (LSTM) [156], which are discussed in the following sections.

3.8.3 Long Short-Term Memory Network

As a remedy to the problem of the vanishing gradient, two German researchers (Juegen Shmidhuber and Sep Hochreiter) proposed Long Short-term Memory Units (LSTMs), a recurrent neural network variant, in mid-1990s. The LSTMs preserve and addresses the error that is generally backpropagated through layers and time [157]. These networks can learn over many time stamps by managing error constantly; addressing the issue of delayed and sparse signals.

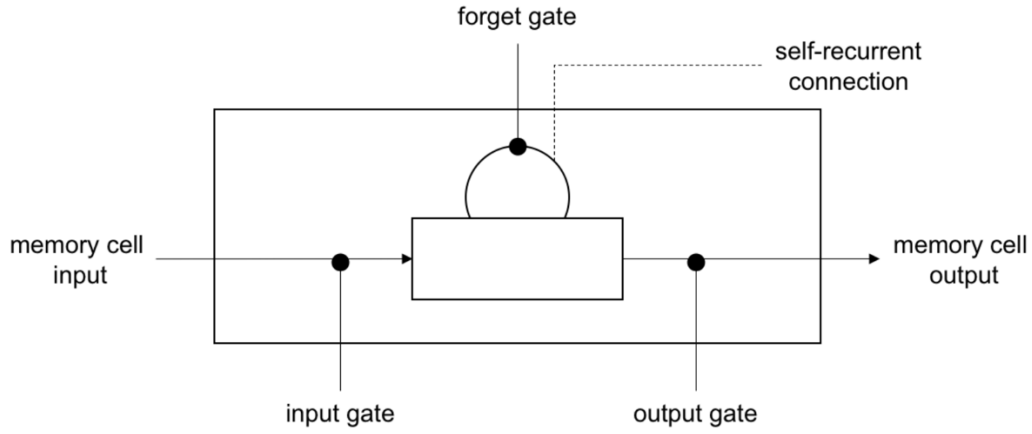

Figure 3.7 Long short-term memory (LSTM) architecture

A gated cell is utilized in LSTMs to capture information that lies outside the reach of the normal flow of RNNs. Like the data in computer memory, the information in gated cell can be written, read and stored. The cell behind the gates, that close and open, makes decisions related to the storage, writing, reading and erasing of information. However, in contrast to digital storage,

these gates store analog signals, implementing sigmoid that fall in the range of 0-1. Figure 3.7 presents the architecture of a LSTM network.

Analog signals are more suitable than the digitals for backpropagation as they are differentiable. The gates used in the LSTMs act upon signals that they receive to pass or block information based on its merits. The information imported is then filtered using weights. These weights are adjusted via the learning process of the networks. Therefore, the cell behind the gates learns about when to leave data, allow to enter, or delete it using an iterative procedure of backpropagating errors, guessing and weight adjustment using gradient descent. Distinctive weights filter information for forgetting output and input. The gate responsible for forgetting is denoted as a linear function and, when open, the present state of the cell is multiplied by the number one and propagated forward to the next time stamp. Furthermore, including '1' as a bias to the forget gate of the LSTM cell has also shown to improve performance [158]. The next step is to decide what values to update (values between 0-1). The number '0' represents unimportant information and the number '1' represents the information is important. The sigmoid function, based on the output it gets, decides what information to keep and what to discard.

Once the work of the input gate is accomplished, the next step is to calculate the cell state. The cell state is then multiplied to the forget matrix, leading to the possibility of value dropping, if gets multiplied by values close to the number

'0'. The output from the input gate is then taken for pointwise addition, and the cell sate is updated to latest values that are relevant to the network. The succeeding hidden state is decided by the output gate.

One may wonder about the existence of the forget gate in LSTMs as the sole purpose of these networks is to link the historic information/input to the current state of information, for performance improvement. However, in some instances it is beneficial to forget some information and set the memory cell to zero again if the text being processed is almost at the end and may not have any relationship with the subsequent documents. Input gates are used to update a cell's state. It does this by passing the current input together with the previous hidden state to the sigmoid function.

3.8.4 Gated Recurrent Units

As in LSTMs, Gated Recurrent Units (GRUs) [155] provide a gating mechanism for RNNs, however they do not have an output gate. This enables LSTMs to write all the content from memory cells to a larger net at every time stamp. RNNs, in general, have a vanishing gradient problem that GRUs aim to resolve. The problem of vanishing gradients is a typical machine learning concern and relates to the size of gradients becoming significantly small or almost vanishing; preventing weights from updating their values, thus impacting learning and prediction performance. Furthermore, GRUs have proven to give better results when working with a smaller dataset. Similar to LSTMs, GRUs are a distinctive

variant of RNNs and possess a similar design, and in some instances produce comparable results and outputs. Figure 3.8 presents the architecture of a GRU.

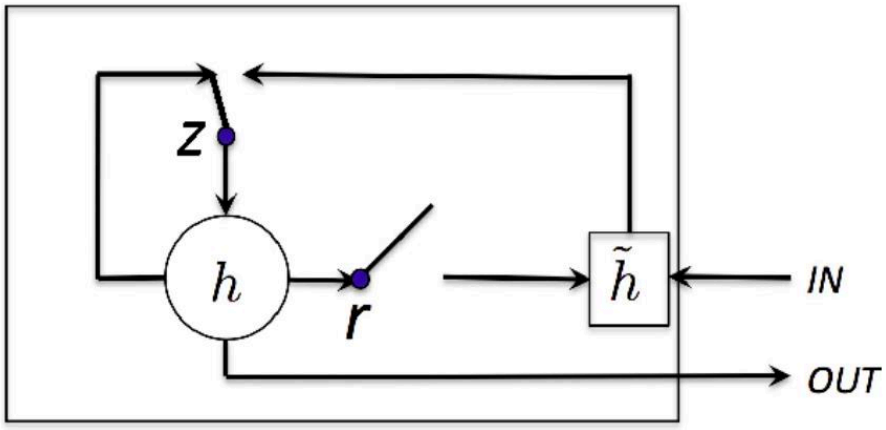

Figure 3.8 Gated recurrent unit architecture

An important difference between the basic RNN architecture and GRUs, is that GRUs support gating within a hidden state, providing a mechanism to update the hidden state when deemed appropriate [155]. The aforementioned mechanism is self-learned by the network to address the short comings of RNNs. By utilizing a rest gate and an update gate, GRUs are well versed in addressing the issue of vanishing gradients that persists in RNNs. A typical reset gate in GRUs regulates the flow of information flowing out of memory and a typical update gates regulates the information that flows in. The reset gate and the update gate are the two vectors that determine which information gets through to the output. These gates can be instructed to remove irrelevant information or to keep the crucial past information for making predictions.

In regard to the architecture of GRUs, there are further several variants with distinctive combinations of bias and gating done utilizing preceding the hidden state. Two commonly used variants are: fully gated recurrent units and the minimal gated recurrent units. These two are analogous in several ways but the reset gate and the updates gate vectors are amalgamated into forget gates [155]. Some of the applications in which GRUs have demonstrated superior results to RNNs and its other variant LSTMs are in handwriting recognition, speech signal modeling, polyphonic music modeling and other natural language processing tasks.

3.9 Chapter Summary

This chapter has provided an overview of technologies implemented in this research. Machine learning and deep learning have been extensively experimented with (in this research) for all the classification tasks performed. The chapter provides brief insights into the inner working of machine leaning and deep learning algorithms, how they are different from each other, and how the differences impact their performance in a broad sense. Machine learning as a technique has been around for a while, however the deep learning technology is a relatively new phenomenon and is evolving. The two families of deep learning architectures, namely CNNs and RNNs, along with their two variants, LSTMs and GRUs, are explained with references. The use of deep learning in behaviour studies is in its nascent form and this research is one of the first to have implemented these technologies to study behaviour patterns in general,

and online ASB in particular. The brief summary of algorithms discusses their inner workings, and how these can be leveraged for social behaviour studies.

CHAPTER 4

INVESTIGATION OF SOCIAL BEHAVIOUR PATTERNS

LBSNs such as Swarm provide a rich source of information on urban functions and city dynamics. Users voluntarily check in to places they visit using a mobile application. Analysis of check-in data offers insights into a user's mobility and behaviour patterns. In this chapter, location sharing data from Swarm has been experimented with to explore spatio-temporal, geo-temporal and human behaviour patterns Sentiment analysis on check-in data using MeaningCloud, psychometric analysis with LIWC15, and descriptive statistical analysis with SPSS was performed to discover meaningful trends and obtain a deeper understanding of human behaviour and mobility patterns. The results showed expressions of people, both negative and positive, when visiting different types of places. Furthermore, the patterns are different for different days of a week as well as for different times of a day, but are not necessarily influenced by weather.

4.1 Introduction

Intra-urban mobility has always been a topic of interest across research communities including urban planning, computer science, physics and geography [159]. Urban planners are interested in improving transport efficiency by investigating spatial and temporal differences of travel time and travel flow, geographers are usually interested in the spatial distribution of

intra-urban mobility, and computer scientists and physicists are interested in modelling the distribution of travel distances in mathematical ways [160-162]. Similarly, human behaviour related to intra-urban mobility has been an area of interests for social researchers. Intra-urban mobility patterns can give insight into behaviour traits for a group of people in a city or in a particular area of a city [163, 164]. The types of places people visit and the time of the day they visit those places explains a lot about their behaviour traits and mobility patterns.

The aim of this chapter is to find new ways to explore human behaviour in today's world of big data. We are generating data at an unprecedented rate and social media has an enormous role to play in it [165, 166]. Every activity that we perform on social media creates data. It has become a part of our daily lives and people use it for various reasons. Facebook, for example, is used for getting news and to stay in touch with family and friends. Instagram and Flickr give users the ability to share photos. Yelp and Foursquare are used to share reviews on services and venues. People use Quora and Reddit as knowledge sharing tools, Pinterest to keep track of things they like, and Swarm (a Foursquare subsidiary) to share locations.

There is no limit to the number of uses for social media sites, and different people use them for different reasons [167, 168]. The use of social media services is on the rise, and they are being integrated into other applications. For example, TV channels are giving their viewers the ability to discuss a particular show or

a sports event live through their own apps and the apps created by others [169, 170]. People are indulging in different social media sites for pleasure and education, and sometimes in response to peer pressure because everyone else around them is using them. All these activities generate data at an unprecedented rate, and this data can be explored for research [171]. This research is the first step towards building a new data mining methodology utilizing user-generated social media data to study human behaviour traits.

Data obtained from online platforms has been used to some extent in business application development and research. Online platforms have utilized this sort of data to improve online customer experience and marketing efforts. Techniques such as customer segmentation have benefited from the use of social media data. However, such data has begun making its way into academic research. Studies have shown that social media data has not been fully explored in academia [172]. Most of the research, specifically in behaviour science, relies on traditional methods of data collection such as interviews, questionnaires and surveys. These traditional methods of obtaining data for human behaviour and mobility patterns research have become inadequate to meet contemporary policy demands [173]. Data obtained via social media platforms and analysed by machine learning algorithms can make the research process more efficient and help us discover patterns that are otherwise difficult to recognize [174-177]. This chapter will use check-in data from Swarm, which is a LSBN. LSBNs have become important sources of volunteered geographic information (VGI). They

can be categorized into two broad categories, 'purpose-driven' and 'social driven'.

A purpose-built network is an application through which people explicitly request another person's location. Some examples of such services are AT&T Family Map and Verizon Family Locator. Whereas a social-driven network is an application through which people broadcast their location to family, friends and followers on the network. Some examples of social-driven networks are Facebook Places, Gowalla and Swarm [178]. Swarm (previously known as Foursquare) is one of the most popular social-driven LBSN apps with the largest number of users and daily check-ins, and has been chosen for this research. The service was initially launched in 2009 as Foursquare and became very popular in a short period of time. The service has more than 60 million registered users and more than 50 million monthly active users. As of October 2020, the platform had surpassed 14 billion check-ins and had reached an average of 9 million daily check-ins [179]. The service has an app for IOS and Android devices and utilizes the device's built-in GPS to track the exact location of user check-ins.

When it first started, Foursquare was a mobile application that provided local search and discovery services along with check-in features. It provided recommendations for places to visit near the user's location. It also worked as a search engine to find local places of interest. The app allowed users to leave tips for different places they checked into. These tips worked as recommendation for

others using the app. In May 2014 Foursquare split into two different apps – Foursquare and Swarm. Swarm is mainly used for check-ins and to keep track of places that users have visited, and Foursquare is used for tips and recommendations for places of interests [180]. Apart from the two most common reasons to use the app, a lot of people use it for:

- keeping track of places visited
- discovering new places
- finding tips about a particular venue
- getting discounts and special offers
- coordinating with friends
- keeping in touch with friends
- letting friends know when a user is available to hangout
- making new friends
- occupying themselves when they are out [167].

Some people also use it due to social pressure because all their friends use it and some use it as a game to play when they are alone [176]. Although different people use the app for different reasons, the underlying attraction is the social aspect of the app [181].

In this chapter, user check-ins are analysed to Identify various mobility and behaviour patterns. Various categories of places people visit the most, busy times to check-in during a day, busy days of the week and sentiment analysis of the messages that users post along with check-in are explored in the chapter.

Patterns based on gender and weather are also explored. The information obtained from the results of this study can be utilized by local government authorities and businesses to plan travel and business activities. The findings can also be utilized to develop behaviour and mobility user profiles which can lead to the development of a targeted recommendation system and lay the ground for other human behaviour related research.

The chapter is organized into five sections. The current section provides an introduction to the overall chapter. It covers aims, along with research conducted. It also covers basic information related to social networks in general, and LBSNs such as 'Swarm' in particular. Section 2 provides insights into the previously published work related to social networks and behaviour studies. It explains how social media data in general and location based social media data in particular, have been used to solve real-world problem. Such data has been used in some research areas, however it deserves more consideration in others. Section 2 also advocates the use of location-based data, not only as an alternative, but as one of the main sources of research data. Section 3 delves deep into the methodology used in this study, covering methods used for data collection, data processing, data transformation and data analysis. Three main types of analysis have been carried out in this chapter and these are explained in detail in this section. Section 4 discusses the results obtained from the study's experimentation. Finally, Section 5 summarizes the findings and presents a

conclusion drawn from them. This section also proposes some of the research work that can be carried out based on these findings.

The main contribution of the chapter is to confirm that different human behaviour patterns can be successfully depicted from individual's online activities. Three primary types of data analyses are performed. Within these, eight other secondary analyses are conducted. Results and findings from these analyses and experiments using MeaningCloud and LIWC15 lays the foundation for experiment conducted and mathematical model developed in the following chapters.

4.2 Related Work

4.2.1 Data from Social Media and Traditional Sources

Large-scale location-based data, when compared to other sources of data from questionnaire and surveys, has a lot of advantages as an indicator of human activity categories, especially when it comes to analyzing trends in entertainment, shopping, travelling and dinning. It provides fine-grained resolution and is readily available. However, it comes with some limitations in terms of how it represents human mobility behaviour [182]. For example, it can contain a group, venue category or age bias that may lead to mobility patterns with certain characteristic. The data may not be suitable for discovering some kinds of mobility patterns, but is still a very valuable source for most research

studies, especially those that study spatial interactions between different categories of places.

Since spatial interactions are measured via human mobility patterns, check-in data from sources such as Swarm has a lot of potential for discovering spatial interactions; the research area of interest for geographers. Geographers can study consecutive check-ins, in a time span shorter than eight hours to ascertain the strength of a relationship between two or more categories of places. A two-way trip between two places strengthens this pattern and shows a stronger relationship. Similarly, more trips between two places or categories shows that these two are spatially connected [182]. Discovering such patterns using data collected from surveys and questionnaires is quite a cumbersome, time consuming and expensive process. LBSN data is more suitable for these and many other related research problems. Researchers from some domains have utilized social media data to some extent, however, it has not been completely explored by others. Zhang et al. [183] studied 13,619 Facebook users to show a correlation between their profile information and the online purchases they made on eBay. The research used various machine-learning models to predict (with statistical significance) the product category from which users would make purchases. The researchers proposed this model to build a 'cold start recommender system.

A cold start recommender system is the only practical solution when an e-commerce company does not have enough information about a user and his purchase history to make recommendation for new products. Most of the recommender systems are built on two widely used techniques: collaborative filtering methods and content-based methods. The collaborative filtering method works on an assumption that users with similar profiles (gender, age group, demographic and ethnicity) have similar characteristics and will most likely buy similar sort of products. The system works well if the seller has enough information about the user. A content-based method, on the other hand, uses information from the web (blogs, review sites, online opinions, tweets and posts) to rank products and recommend those to buyers. This approach may work to some extend but not always. A typical recommender system requires a lot of user information, collected through a user's web and purchase history. This information, collected through traditional methods, is not available in this scenario to make a cold start recommender system but data from social media fills in the gap very well.

The model that Zhang et al. [183] suggested in their study has another way to look at the same problem, i.e. to recommend the most appropriate products to buyers. E-commerce sites such as Amazon and eBay encourage buyers to share their purchases on social networks. Buyers who share their purchases on social media, by visiting sites via links provided on an e-commerce seller's page, end up sharing some of their profile information such as pages they like and places

they have checked into, with the ecommerce seller. The places that a user has checked into the past tell a lot about their taste and preferences. For example, if a person has checked-in at a fashion store, the ecommerce site (Amazon or Ebay) can recommend fashion accessories to the user on the same or the next visit the user makes to the site. Similarly, if a user had checked into a bookstore in the past, the chances are that if the ecommerce site offers a newly released book that is of interest to the user, the user might look into it.

The pages that a user likes on their social media profile also provide a lot of information about the things the user may buy. For example, if a user has liked a page of a band/music genre, the seller can recommend music from the same band or genre to the buyer. Similarly, if a user has liked the page of a sports team, the seller can try to sell tickets for that team's next game. So, unlike a collaborative filter or a content-based system, a cold start recommender system that can be fed with information from a user's check-in history, can work well for recommending items to a new buyer whose buying history is not available to an ecommerce site.

4.2.2 Real world applications of Location Based Data

Aliandu [61] study used data from Foursquare to rank popular shopping, accommodation and culinary locations in Kupang city, Indonesia with 77.48% accuracy. She extracted text data from Foursquare using an API and performed sentiment analysis to find positive, negative and neutral sentiments for tourist locations in Kupang. Her study shows how social media data can be used to

discover popular locations in a city and provide opportunities to discover other patterns of human mobility behaviour. The study suggested that such data could be very useful for studying spatial and temporal trends of human mobility.

Zhou et al. [63] showed how check-in data from Foursquare and Twitter can be mined to discover urban functions such as transport needs, business distribution and development trends. They showed how using check-in data over traditional sources of data such as photography, observations, cognitive maps and remotely sensed images made more sense. They collected 70 million geo-referenced tweets across the US. Using geo-tagging, they filtered 664,705 and 911,301 tweets for Boston and Chicago, respectively. These tweets were then used for both large-scale and fine-grained analysis to detect hotspots and citywide dynamics in both cities. City authorities can use this intelligence for future planning and development.

Similarly, Hasan et al. [62] used large-scale location based data from Twitter and Foursquare to discover urban human activity and mobility patterns in three different cities: New York, Chicago and Los Angeles. They extracted check-in data that was initially created on Foursquare and shared via Twitter. The check-in data indicated the types of places people visited based on ten location categories developed by Foursquare. Every check-in was associated with a

different category, and by analyzing these the authors discovered mobility patterns in New York, Chicago and Los Angeles.

Noulas et al. [184] used check-in data to present the spatio-temporal patterns of users' activities and conduct place transition analysis. They collected Foursquare check-in data from two different sources to avoid any platform bias and found the same patterns from both sets of data. They collected 12 million check-ins related to different venues via Twitter, and requested a large dataset of check-ins directly from Foursquare. Similar trends were found from both datasets; validating check-in data as an important source of research data.

Check-in data can be a very good avenue to discover collective user activities. A higher number of check-ins at a particular bar in a city can imply that the bar is a popular venue. Similarly, a large number of check-ins at a train station can indicate that it a busy station compared to some of the others in the area. Authors in [184] used a cumulative distribution function to discover top ten places for check-ins, both during weekdays and weekends. The study not only discovered patterns in how people moved around during weekdays and weekends, but also found similar patterns for different hours of the day. Most of the findings were in line with previous research findings in the area that used different, large and cumbersome sources of data such as surveys and questionnaires. The top place for check-in during week days was train station followed by supermarket, bar, mall, highway, café, office and home. Unsurprisingly, the pattern was different

for the weekend with hotel/bar replacing train station as the top activity, followed by restaurant, department store, supermarket, highway, café and home. Office was not in the list of top 10 check-ins during the weekend, and its spot was filled by department store. Supermarket was ranked higher during the weekend than weekdays. This is in line with common knowledge that people visit department stores during the weekend and most people do their grocery shopping during the weekend as well.

Getting these sorts of results from check-in data is promising as they validate the significance of such data for research. Since check-in as a data source is relatively new, these results show huge potential. The results also show patterns such as high check-ins numbers in the morning when a lot of people are at train stations and/or going to work. Check-ins increase during lunchtime when people are out visiting cafés and restaurants for meals. Check-ins peak in the evening when most people finish work and are at places such as train stations, restaurants, bars, highways, supermarkets and homes. The patterns discovered were from a large data set and related well to our daily activities.

The study also analysed a different sort of relationship between the places people visited. About 10% of users logged two check-ins within a short period of time (less than 60 minutes). These two temporal adjacent check-ins by the same user can signal an important correlation between the two venue types that the user has visited and can indicate a temporal relationship. Similarly, two

check-ins made by a single user at two different venues, that are not too far away from each other (less than 1 km) can indicate some sort of spatial relationship between the two places. It is possible that both venues fall into the same category such as bar or restaurant. These can also be from different but related categories of venues such as movies and restaurants. Finding one of these venues can lead to predicting a check-in at the next related venue [185].

LBSNs can also be useful to solve a cold start location prediction problem as demonstrated by Gao et al. [186]. Their study proposed a 'Geo-Social Correlations" model that can assist in predicting a user's next check-in. When a user checks into a venue, there are two possibilities: either the user has visited that venue before, or it is a new venue. The choice of new venue may depend on the user's visit history or the visit history of the user's friend. A check-in can also be correlated to check-ins by other people in the same area, who may or may not be user's friends. The probability of predicting a next check-in using this model was significantly higher than a random guess based on users' past check-in histories. The study also found that a new check-in by a user was more strongly correlated to check-ins by other people in the area than It was with the user's friends. There is some correlation between a user's and his/her friend's new check-in, but it is not as strong as the correlation between a user's check-ins and the check-ins of other people in the same area. An example of this would be a check–in at a local neighbourhood café or a local train station where two people from the same community are more likely to check-in even though they are not

friends with each other. This study [186] is one of the first in the area to come up with such a model to predict users' next check-ins.

The application of such a model could be to a recommender system for targeted marketing. A recommender system, as a typical business application, has been discussed in this chapter however, it is not the only business application for social media data. Since the data has prediction potential, it can and has been used in financial markets as an alternative source of data [58, 187-189]. Organizations such as hedge funds and investment banks use check-in data to discover various trends in a geographical region to assist them with investment decisions. These organizations discover knowledge from such data in three different ways. First, they collect the data and use data mining techniques in-house to discover trends. Second, they buy data from Foursquare and have it processed (cleaned and mined) by a third party. Third, they buy data instead of collecting it themselves, and do all the processing and knowledge discovery in-house. Foursquare offers a service called "Place Insight" through which businesses can buy data directly from it and use it for analysis [190].

To conclude this section, online social activities of a person are indicators of his/her real world behaviour [6, 52]. These activities can reveal the types of places people visit and the things they do [61-63]. Such activities can be studied to discover the behaviour patterns of individuals and groups of people [56]. Furthermore, weather can influence the types of activities people engage in, and

the behaviours they exhibit [64]. Based on these theories, the chapter will further explore human behaviour and mobility patterns and how these are impacted by weather.

4.3 Methodology

4.3.1 Data Source

The data source for this research project is check-in data from Swarm (formally known as Foursquare). The app allows users to check-in when they are at a particular venue. Users launch Swarm and log a check-in. Once logged in, the check-in details are shared with the user's connections on the platform. Check-in data is classified as large-scale location-based data and contains mainly text along with time and date data. Typical check-in data contains the following fields:

- user name
- message, if any, along with the check-in
- time of the day, day and date
- category of the venue
- check-in location (latitude and longitude coordinates)
- gender.

Most users of Swarm check-in at a venue using the mobile app, and share that check-in via Twitter. Then the check-in is broadcast to the user's followers on the Twitter. Tweets containing Swarm check-ins are the data source for this project

4.3.2 Data Collection

Data for this study was collected from Twitter and contains Swarm check-ins shared on Twitter. Collecting data directly from Swarm for research purpose involves privacy issues and only those check-ins publicly shared on Twitter, were collected. Data for this study was collected manually unlike most other similar studies that use Twitter API or some sort of software tool to collect Twitter data. Some researchers have used web scraping to collect data from Twitter. The limitation of using a Twitter API, web scrapping or software tools to collect data from its platform is that one can only go back a week at a time to access the data. Twitter does not allow users to access data that is more than one week old. Almost all software tools available for collecting Twitter data, have its API working in the background and so are dictated by the same Twitter API protocol.

As this study needed data from the summer and winter of 2017 and needed to collect that data freely, manual data collection was the only option available. Another limitation of using web scraping, Twitter API or a software tool to collect data from Twitter is that Twitter only allows a random 1% of its data to be collected with these methods. Using a manual method, we were able to collect all the check-in tweets created in Melbourne metropolitan area. Two 30-day time periods were selected, one from summer 2016-2017 and another from winter 2017. For winter, the data was collected from 1st June 2017 to 30th June 2017, and for summer it was collected from 15th January 2017 to 13th February 2017. To

measure any differences in behaviour patterns due to warm or cool weather, data was collected from a summer month and a winter month. To analyze the mobility patterns in Melbourne, data from the summer and winter months were combined into one data set.

June is the first month of winter in Australia and we chose this month to collect our winter data. We could have chosen the month of December to collect our summer data (as December is the first month of summer in Australia), however, we wanted to avoid any bias due to the holiday period and chose a different time period. December and first two weeks of January make up the holiday and celebration period and a lot more people visit bars, restaurants and other places of entertainment during this time. To avoid any bias in the dataset, we chose a 30-day period starting from 15th January, when most people are back to their normal lives after New Year celebrations and holidays.

To collect Swarm check-in data through Twitter, Twitter's advance search tool was used to select the check-ins from the Melbourne area for the desired dates. Each check-in was manually copied from Twitter and pasted into an Excel spreadsheet for further processing. A total of 1499 check-In tweets were collected from the 30-day period in the winter, and a total of 1833 check-in tweets were collected from the 30 day period in summer. These are all the check-in tweets that were created on Swarm, in Melbourne, and shared on Twitter

during those two 30-day periods. We could not have done this by any other data collection method.

4.3.3 Data Pre-Processing

Swarm assigns every check-in by user to a venue category. It has 10 top-level categories and these are divided into second level categories (subcategories). Second level categories are further divided into third level categories, and so on. For example, 'Art and Entertainment' is one of the top-level categories. It has sub (second level) categories such as Amphitheater, Arcade, Art Gallery, Memorial Site, and Movie Theater etc. Movie Theater has third level categories such as Indie Movie Theater, and Multiplex. Similarly, 'Food is a top-level category which has 'Asian Restaurant' as a second level category, which has a 'Chinese Restaurant' as a third level category, which has a 'Hakka Restaurant' as a fourth level Category. The maximum level of category depth in Swarm is four. Almost all check-ins belong to some first level category and then, based on the depth of subcategories, may belong to a second, third or fourth level category as well. The top ten (first level) categories in Swarm are:

1. Art & Entertainment
2. College & University
3. Event
4. Food
5. Nightlife Spot
6. Outdoor Recreation

7. Professional & Other Places
8. Residence
9. Shop & Service
10. Travel & Transport

For this chapter, considering the nature and scope, more appropriate venue categories were developed. The size of the top-level category was increased to 20 to cover the breadth and to have a better understanding. Depth was reduced from 4 to 2. This was done for simplicity and to streamline venue categories. All check-ins collected belonged to a second level and a first level category. The following are the venue categories and some of their subcategories used in this study:

1. Airport
2. Bank – ATM, Bank
3. Bar - Pub, Bar, Night Club
4. Education - College, School, University, Library
5. Entertainment - Show, Circus, Adventure Sports, Movies
6. Food - Restaurant, Cafe
7. Grocery - Supermarket, Petrol Station, Milk bar
8. Gym - Gym, Yoga Studio, Aerobics, Spa, Sauna
9. Home - Home, Apartment, Unit
10. Hotel - Hotel, Hostel, Motel
11. Landmark - Places of Interest in Melbourne

12. Medical - Doctor, Hospital, Pathology, Pharmacy
13. Neighborhood - Suburb, CBD
14. Outdoor - Park, Beach, Hiking
15. Public Transport - Train Station, Bus Stop, Tram Stop
16. Religious Place - Church, Temple, Mosque, Synagogue
17. Salon - Hair Salon, Beauty Salon, Massage Parlor
18. Shopping Centre - Neighborhood Shop, Shopping Center, Shopping Strip
19. Sports - Stadium, Community Sports Club, Any other Sport Complex
20. Work - Office Building, Corporate Office, Other workplace

4.3.4 Data Transformation

Once the check-in tweets were collected from Twitter, they were cleaned and transformed into a form that could be analyzed easily. Each of the venue categories was assigned a number from 1 to 20 as shown in the above list. For example, 1 represented a check-in at an airport, 2 represented a check-in at a bank, and 3 represented a check-in at a bar, and so on so forth. Similarly, each day of the week was also assigned a number ranging from 1 to 7 (Monday – 1, Tuesday – 2, Wednesday – 3, Thursday – 4, Friday – 5, Saturday – 6, Sunday – 7).

For gender classification, males were assigned a value of 1, female 2 and couple 3 (a joint check-in by both a male and a female). Days from both 30-day periods were assigned values ranging from 1 to 30. So, for the winter month of June, 1

represented 1st June 2017, 2 represented 2nd June 2017, and so on so forth. For summer, we started collecting data from the middle of January, so 1 represented 15th January 2017, 2 represented 16th of January, and so on so forth.

Transforming time data was a bit tricky because it was in a 12-hour format (AM and PM) when we collected the tweets. So, the first step was to convert them to a 24-hour format (e.g., 13:30:15). Once in a 24-hour format, we divided the 24-hour day into four parts – morning, afternoon, evening and night. For convenience, we could have divided the day into four six-hour slots but that would not have worked for this study, so we decided to classify timeslots as follows:

- Morning (5am – 12pm)
- Afternoon (12pm – 5pm)
- Evening (5pm – 9pm)
- Night (9pm - 5am)

Keeping the night slot for eight hours made sense because most people in Melbourne are sleeping during those hours. The shortest slot was for the evening (four hours) and the morning and afternoon timeslots were allocated seven and five hours respectively. Once this was done, each timeslot was allocated a number from 1 to 4. 1 represented a check-in logged in the morning between 5am to 12pm, 2 represented afternoons, 3 represented evenings, and 4 represented check-ins made during the night. Text messages within the tweets

contained emojis and other symbols. These were kept unchanged to address any emotion expressed when conducting psychometric analysis.

4.3.5 Data Analysis

In this chapter, three primary types of data analysis were performed. Within these, other secondary analyses were also performed as follows:

1. Text analysis
    i. Language analysis
    ii. Topic analysis
    iii. Text clustering analysis
    iv. Sentiment analysis
2. Psychometric analysis
3. Statistical analysis
    i. Spatio-temporal user activity analysis
    ii. Gender-based activity analysis
    iii. Geo-temporal patterns
    iv. Check-in dynamics

Analyses and results for these techniques are discussed in the following sections.

4.3.5.1 Text Analysis:

Text analyses were done using the MeaningCloud software application. A lot of people, when checking in at different venues, also like to express their feelings

and thoughts. They write small notes expressing their sentiments, emotions and any other thoughts they have at the time of check-in. An example of a short message posted alongside a check-in is:

" Laura @ViataDulce
This mall has changed so much. Starbucks and Grilld for dinner. #Eastland (@ Eastland Shopping Centre)"

These messages can offer a meaningful insight into user behaviour and emotions. Different techniques can be used to analyze these text messages and in this chapter MeaningCloud was used for such analyses. MeaningCloud (previously Textalytics) is a software application that enables text analytics and semantic processing of text data. The tool is offered on the cloud, as software as a service (SAAS), and on premise mode [191]. Results of these analyses are discussed in section 4.4 'Results and Discussion'.

4.3.5.1.1 Language Identification with MeaningCloud

This feature identifies the language in which a document is written when the document is fed into the software tool. However, in this experiment, the different tweets related to separate check-ins were captured in an Excel spreadsheet. These were fed into the software application, with each cell of the sheet represented as a separate entity. Analyses were done on both summer and winter data by combining these into one dataset. MeaningCloud searched each cell and identified the language in which the text message was written. It displayed an output (name of the language) in the cell adjacent to the cell containing the message.

4.3.5.1.2 Topic Extraction

MeaningCloud's text extraction feature returns a main theme/topic from the text. The software goes through each and every word of a phrase and analyses semantics from the text, relating syntactic structures from the level of sentences, phrases, clauses and paragraphs in relation to the entire text and language-independent meanings. Basically, the software tries to establish relationships between the different words of a sentence to gauge the meaning that these words convey. The next step in this analysis is to establish the meaning of each and every sentence relative to the whole text and identify a common theme for the text. In this case, since text data is in different cells, the software traversed through each and every cell and produced the main theme from each cell.

4.3.5.1.3 Text Clustering

A text cluster forms a cluster of the most common words in a document. The MeaningCloud software application identifies all the words in a text and counts the number of times these appear in a document. These are then presented in ascending order as an output. The most common words used in a document have the highest counts and the least used have the least counts. Text cluster was performed separately on data from the summer and winter periods. It was conducted to identify the most popular words used during both winter and summer seasons and to understand what people talk about depending on the weather at the time of the year.

4.3.5.1.4 Sentiment Analysis

Humans are emotional creatures and like to express their emotions when visiting different places. Sentiment analyses were conducted to discover emotions that people express at different places depending on the weather. This experiment was also done to gauge the percentage of people who express some sort of emotion when logging a check-in. The summer and the winter data was analyzed separately to observe any difference in the way people express emotions depending on the weather at the time. Sentiment Analysis is one of the most popular analyses for which MeaningCloudis used. The software analyzed each and every tweet and assigned them to one of the five sentiment categories – Strong Negative, Negative, Neutral, Positive and Strong, and Positive. These were then aggregated to analyze all the emotions expressed.

4.3.5.2 Psychometric Analysis

Psychometric analysis on check-in data was performed to discover emotions not covered by the five sentiments discussed in the previous section. Linguistic Inquiry and Work Count (LIWC2015) were used to perform these analyses. The tool was developed by James W. Pennebaker, Ryan L. Boyd, Kayla Jordon, and Kate Blackburn of The University of Texas, Austin. Every day words that people use can provide rich information about their fears, anxiety, anger and beliefs [192]. The application can be utilized to study various cognitive, emotional and structural components present in an individual's written text sample and verbal speech.

The software consists of a dictionary with more than 6400 words, word stems and emotions. These words are associated with one or more word categories and sub-dictionaries. One example of the dictionary structure is that the word 'cry' is associated with five different categories – Negative emotions, Sadness, Overall affect, Past focus and Verbs. All sadness words belong to a broader Negative emotions and Overall affect categories and so on so forth. For each text file (or a cell in the Excel spreadsheet) the software outputs 90 different variables. These include word count, language summary, general description, linguistic dimensions, psychological construct, word categories, punctuation categories and language markers [193]. The check-in data was analysed for three main psychometric properties: anxiety, anger and sadness. Considering their scope, these three psychometric analyses are the most relevant to the experiments in this chapter.

4.3.5.3 Statistical Analysis

Descriptive statistical analyses were performed on the data set to explore mobility and behaviour patterns. Analyses were done to explore how people moved around in Melbourne city based on the time of day and the day of week. Different venue categories were explored to see which ones were the busiest at different time frames. Behaviour and mobility patterns based on gender were also analyzed to see the difference in places popular among males and females. The analyses looked into different times of the day and different days of a typical

week to gauge differences in behaviour patterns between male and female users. The results are discussed in the following section.

4.4 Results and Discussions

4.4.1 Text Analysis

4.4.1.1 Language Analysis

Language analysis was done to see if users use any language other than English in the messages associated with check-ins. Our hypothesis was that even though Melbourne is mainly an English-speaking city, there are still a lot of people who would write comments in different languages. The results show that our null hypothesis is true and the use of languages other than English is quite prominent. Arabic, Chinese, Japanese and Korean were some of the languages used extensively in the comments associated with check-ins. It is not clear whether people posting messages in these languages were natives of Melbourne or were travelling here from other places. It is known that Melbourne is a multicultural city, with people from many different countries around the world settled down here. Therefore, it is quite possible that some of those tweets written in languages other than English were posted by people who live in Melbourne permanently. Melbourne is a multicultural city with people speaking many different languages and the use of different languages in tweets is evidence of this.

4.4.1.2 Topic Extraction

After conducting topic extractions on the comments posted, some of the most common topics related to check-ins, and the categories they belonged to were determined. The results are shown in the Table 4.1. It was discovered that Melbourne was the word most used in all the check-in tweets, followed by 'Tullamarine' (Airport), Shopping, Train, and so on so forth. A lot of check-ins did not contain any extra comment and were only generic such as "I'm at _ _ _ ". Since the data was collected from the Melbourne metropolitan area, it is not surprising to find the city name as the topic of discussion in a lot of check-ins. The trending words are in line with some of the most popular venue categories that people visit and check-in at in Melbourne.

Table 4.1 Topic extraction from check-ins

Trending	**Category**
Melbourne	Neighbourhood
Tullamarine	Airport
Shopping	Shopping
Train	Public Transport
Drink	Bar
Dinner	Food
Game	Sports
Lunch	Food
MCG	Sports
Breakfast	Food

People in Melbourne love sport and that is why it is known as sports capital of Australia. This explains 'Game' and 'MCG' (Melbourne Cricket Ground) in the top ten topics of discussions. Both summer and winter are busy times for sports.

Cricket and tennis tournaments cover most of the summer, and football (AFL) covers most of the autumn and the winter (March to September) seasons. Apart from sport, Melbourne is known as one the best places in the world to eat and drink. There are many places people can visit to have a meal or a drink and enjoy the variety that Melbourne has to offer. The café and bar culture of Melbourne is world-renown. This may be one of the reasons that topics such as 'Drink' and 'Lunch' are among the top-ten topics discussed in text posted.

4.4.1.3 Text Clustering:

Table 4.2 shows the top ten words used in the check-in tweets in Melbourne during the summer 2016-17 and the winter 2017 periods. A separate analysis was conducted on the summer and winter data to see any difference in the words used, implying the places visited. 'Vic', 'Melbourne', and 'Victoria' are the top three words used in all tweets. The explanation for this could be that every check-in has a locality address at the end. Since all tweets were collected from Melbourne, they all have at least two of these three words. As can be seen, the words are quite similar and almost in the same order in both the winter and summer periods, implying that there is no significant different in the type of places people check into and talk about. As it can be seen in Table 4.2, the word 'Outdoor' was used a lot in summer compared to winter. Warm weather brings a lot more people out to get involved in outdoor activities, so it was no surprise to see this word used in summer and not so much during the winter period.

Table 4.2 Text clustering from check-ins

Text Clustering	
Summer	**Winter**
Vic	Vic
Melbourne	Melbourne
Victoria	Victoria
Tullamarine	Tullamarine
South	Station
Station	ShoppingCentre
ShoppingCentre	South
Outdoor	East
Australia	Epping
Coffee	Southbank

Another difference between the summer and winter results is that, in the winter two suburb names appear, but they are missing from the summer result. The results also show the word 'coffee' in summer tweets indicating that more people go out for coffee more in warm weather than in cool weather. 'Shopping Centre' was among the top ten words for both seasons. This may indicate that shopping centres are one of the popular venues. People may like to broadcast this venue category more often than not. This may also imply that shopping is one of the main activities that most people get engaged in for pleasure and fun. Melbourne is home to many shopping centres, including the world-famous shopping centre Chadstone Shopping Centre.

4.4.1.4 Sentiment Analysis

Sentiment analyses were performed separately on the summer and winter datasets to see how people express their emotions when visiting places in these

two different seasons, and whether these emotions are impacted by weather. It was found that sentiments expressed during both seasons were quite similar as opposed to one of the hypotheses that people write positive things (happy text) about places they visit during summer weather more compared to winter weather. Some of the literature in behaviour science suggests that people are happier in summer than in winter, however, those emotions were not clearly obvious from the results. It may very well be possible that, even though people are happier during warm weather, they may not like to express those emotions to everybody through check-ins.

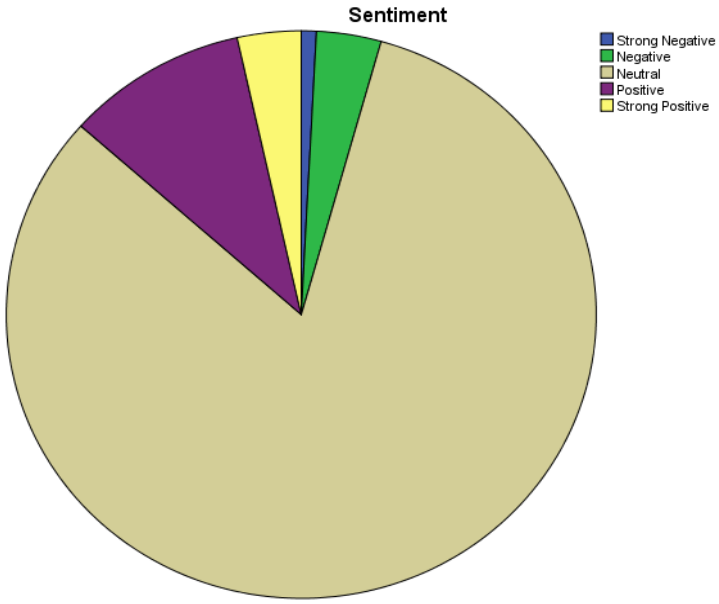

Figure 4.1 Sentiment analysis for the summer

Figures 4.1 and 4.2 show the results of the sentiment analysis for summer 2016-17 and winter 2017 respectively. Each colour represents a different sentiment expressed during each season.

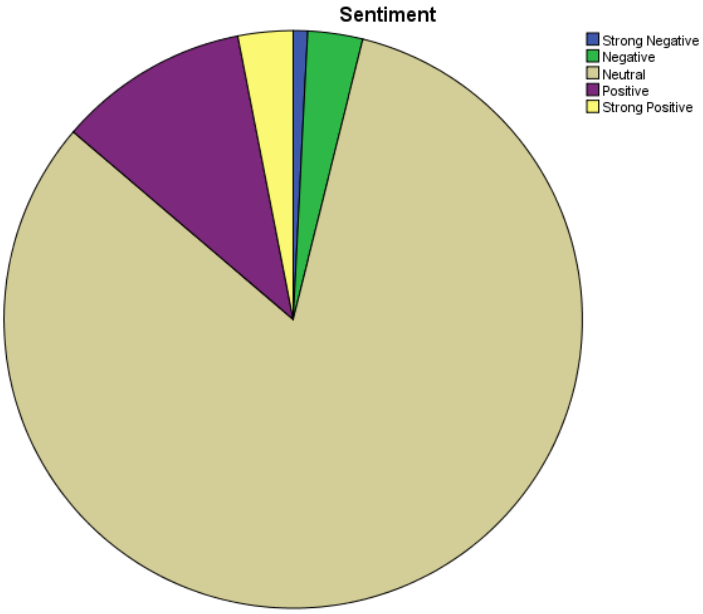

Figure 4.2 Sentiment analysis for winter

The results of our analysis shows that most check-in tweets are neutral in nature, and that means most users in Melbourne have not expressed strong emotions related to places they have visited during those two different time periods. As can be seen from the two pie charts (from the summer and the winter respectively), the percentage of Neutral tweets/check-ins in summer was 82.2%, which is almost identical to 82.4% in winter. Strong Negative tweets in summer were 0.8%, which is exactly the same as the number for winter tweets. Similarly, the percentages for Negative, Positive and Strong positive tweets in summer are 3.5%, 9.9% and 3.5% and in winter are 3.1%, 10.7% and 3.1%. As can be seen from the two figures above, there are no significant differences between the summer and the winter results. People express similar sorts of sentiments for places they visit in both seasons. These numbers may not be true indications of how people actually feel during these two seasons as it is very possible that

people do not like showcase their sentiments openly via social media platforms to their friends and family members; preferring to keep them private.

For further analysis, tweets and associated messages were analysed to investigate the reasons the percentage of Neutral tweets was high compared to the four other categories. It was discovered that a lot of check-ins did not have any comments/text associated with them. Users related to those tweets have just checked-in to some place and shared the location with friends and family without expressing any emotion or sentiment at the time. There are also some tweets in which the text is neither positive nor negative. These check-ins are also tagged as Neutral.

Nearly 18% of the check-in tweets expressed either negative or positive sentiment related to the places that were visited. Almost all the negative tweets made some sort of complaint about the place visited, and almost all the positive tweets have shown some sort of happiness and excitement at visiting venues. The results also show that, in both the seasons, there are more positive tweets than the negative tweets. This may imply that most people feel happy and excited when visiting places. A small number of negative sentiments may be the result of bad service or other negative experiences during their visits to venues. Since the number is minuscule compared to number of positive sentiments expressed, the results show that most people are neutral or happy rather than

unhappy and sad while visiting places or going out. The weather does not seem to have much impact on visitors' sentiments towards the places they visit.

4.4.2 Psychometric Analysis

Psychometric analyses were conducted on the dataset to gauge whether people became anxious, angry or sad when visiting places. Table 4.3 shows some of the examples of messages from the dataset and the psychometric properties they exhibit.

Table 4.3 Psychometric analyses from check-in message 1

Message with Check-in	***Anxious***	***Anger***	***Sad***
Panic!At The Disco (@The Disco in West Melbourne, VIC)	*1*	*0*	*0*
Sam ??Where are you? (@ Tallboy and Moose in Preston, VIC)	*1*	*0*	*0*
Ahhh I'd forgotten how rude the staff are here. Welcome back Robstar!	*0*	*1*	*0*
Where every part of order is wrong. Why? This place general lily sucks (@ The Coffee Club in Airport West, VIC)	*0*	*1*	*0*
olives here suck. maybe they do better gibson martinis? ??(@ Cookie in Melbourne, VIC)	*0*	*1*	*0*
Missed the train by just a few seconds :') WHY IS THERE EVEN AN EXTRA PLATFORM HERE ??(@ Surrey Hills Station)	*0*	*0*	*1*
Last time I used public transport was in Greece. Gosh I miss it	*0*	*0*	*1*
im gonna miss this place ??(@ Grill'd in Cheltenham, VIC)	*0*	*0*	*1*

The '1' in a cell next to a message represents the presence of the related psychometric property and '0' means absence of that property. Data for both the summer and the winter season were analyzed separately to discover any behaviour differences due to the weather. The results are presented in the Table 4.4. It was discovered that the results were quite similar regardless of the

weather. This is contrary to a common belief that people are happier and feel more positive in summer than they do in winter.

Table 4.4 Psychometric analyses from check-in messages 2

	(Summer)		(Winter)	
Anxiety				
	Frequency	Percent	Frequency	Percent
Neutral	1829	99.8	1497	99.9
Anxious	3	0.2	2	0.1
Total	1832	100.0	1499	100.0
Anger				
	Frequency	Percent	Frequency	Percent
Neutral	1813	99.0	1492	99.5
Anger	19	1.0	7	0.5
Total	1832	100.0	1499	100.0
Sad				
	Frequency	Percent	Frequency	Percent
Neutral	1816	99.1	1485	99.1
Sad	16	0.9	14	0.9
Total	1832	100.0	1499	100.0

As discussed, in the previous section, many users do not write/share any comment when checking-in and these check-ins are classified as Neutral messages. As can be seen from the results, most messages were Neutral in both summer and winter. There were a few messages that showed some emotions (Anxiety, Anger and Sadness), however, the number was small. For both summer and winter, the percentage of such messages is less than 1%. Among others, there could be two different explanations for such a low number of check-ins associated with such behaviour. First, users are mostly in good mood,

when visiting places and look forward to having fun, and that is why they are not anxious, angry or sad when going out. Another explanation could be that even if some users get anxious, angry or sad when they are at a particular venue, they probably like to keep their emotions to themselves and do not want to broadcast them to others.

4.4.3 Statistical Analysis

4.4.3.1 Spatio-Temporal User Activity Patterns

In this subsection, the results of the statistical analysis for spatio-temporal activity patterns are presented. The results suggest that activities in Melbourne differ during the course of the day and the week. Meaningful patterns, that are tightly related to human activity from a spatial and temporal point of view, were discovered. Figure 4.3 shows how activities in the city increase as the day proceeds.

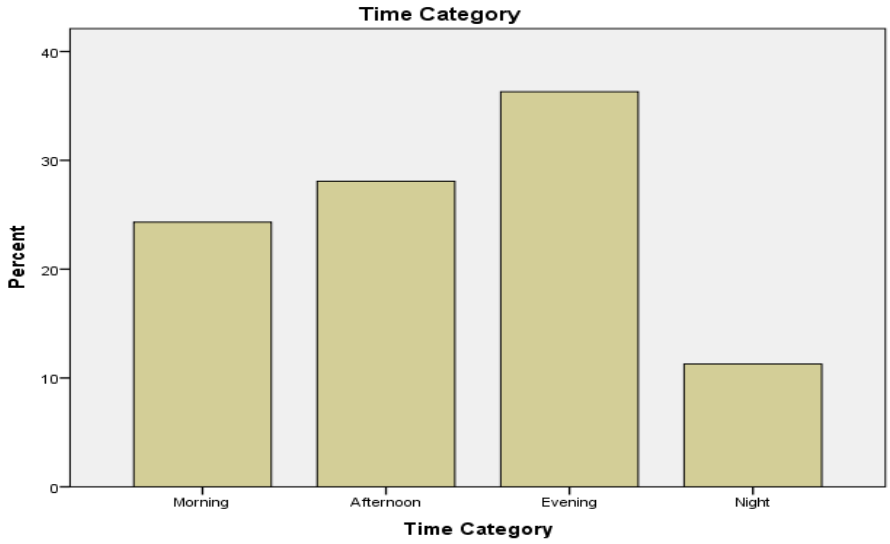

Figure 4.3 Spatio-temporal user activity patterns during a typical day

More people are active and moving around in the afternoon than in the morning and this activity increases in the evening before it starts slowing down after 9 pm. Evening (4 pm to 9 pm) is the busiest time in the city with 36.3% of the check-ins, followed by 28% in the afternoon, 24.3% in the morning and only 11.4% at night. The explanation for evening being the busiest time is that most people finish work between 4 pm and 9 pm and visit places such as train stations, restaurants, bars, sports centres, movies, and airport etc. Most people sleep at night so there is not much activity in the city. This may not have been the case if we had collected data from city such as New York where people are active almost any time of a day. Table 4.5 shows the activities during the day based on category of venue in the city.

It can be seen that the 'Food' category has the highest number of check-ins during the evenings and that is when a lot of people go out for meals. Beside restaurants, a lot of people also visit bars and pubs for after hour drinks. The Table 4.5 shows some activity during the night and most of this is at the Tullamarine Airport, Late nightclubs, and Restaurants. 12.5 % of those check-ins at night are at home or at a hotel. These people are not really moving around the city but are at home. The check-ins at hotels and homes are the highest at nights compared to other time frames, indicating that people are returning back home after finishing their days. Most check-ins in the Morning are at Café's, Airport and Work; with cafés being the busiest. Twenty-four out of 200 check-ins at work

are during nights, indicating that around 12% of the total workforce do some sort of night work/shift. This sort of work is most likely to be at places such as hospitals, nursing homes, hotels, nightclubs etc.

Table 4.5 Spatio-temporal user activity pattern 1

VENUE TYPE		TIME CATEGORY				
		Time Category				
		Morning	Afternoon	Evening	Night	Total
VENUE TYPE	Airport	122	38	74	31	265
	Bank	2	7	4	1	14
	Bar	6	34	116	65	221
	Education	70	52	40	7	169
	Entertainment	12	45	88	23	168
	Food	219	354	426	93	1092
	Grocery	27	35	37	8	107
	Gym	22	5	28	5	60
	Home	5	14	35	23	77
	Hotel	16	6	12	24	58
	Landmark	10	25	23	11	69
	Medical	18	17	16	1	52
	Neighbourhood	13	10	13	16	52
	Outdoor	10	45	30	5	90
	Public Transport	64	46	56	17	183
	Religious Venue	3	5	4	1	13
	Salon	1	7	10	0	18
	Shopping Centre	47	100	73	8	228
	Sports	45	41	95	13	194
	Work	98	49	29	24	200
Total		810	935	1209	376	3330

Gym check-ins are highest in the evenings and mornings, showing that people still prefer to go for exercise in the evening and morning even though most gyms in Melbourne are open 24 hours. Table 4.5 also shows that around 8.3% of gym goers prefer to exercise late at night. Most people visit banks or ATM in the

afternoons. This resonates with people having quick visits to banks and ATMs during their lunch breaks from work. Visits to doctors, hospitals and groceries stores are steady during the day. Public transport is busy in the mornings when most people go to work and evenings when they come back home. The Airport is busiest in the Morning with 46% of total check-ins taking place during this time. This activity indicates that most flights are schedule early in the mornings and in the evenings.

Similar to activity variations during the day, activities also vary during the week. Figure 4.4 shows how mobility and activity increase as the week proceeds. The number of check-ins is highest for Sundays followed by Saturdays and Fridays. Most people do not work during weekends and that enables them to spend time doing what they like the most, such as sports, meals, movies, shopping, etc. and that is why weekends are busier than weekdays when most people have more time to go out and move around the city.

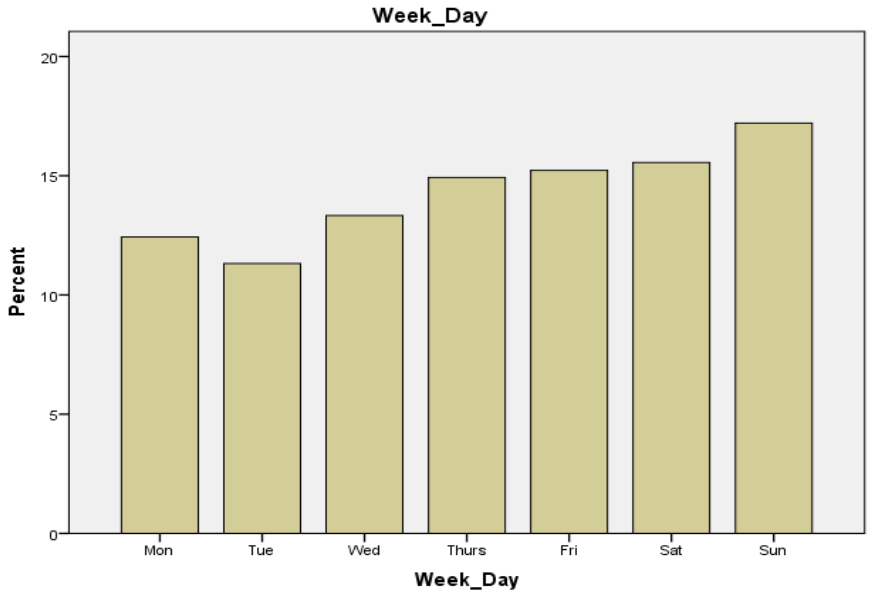

Figure 4.4 Spatio-temporal user activity patterns 1

Tuesdays are the quietest days, followed by Mondays, then as the week passes, people start moving around. One explanation for this could be that during the start of the week, most people focus more on their work and as the week passes, they feel more relaxed and start going out more for different activities. Table 4.6 shows a breakdown of temporal patterns based on different venue categories during a typical week.

Table 4.6 Spatio-temporal activity pattern 2

VENUE TYPE	WEEK DAYS							
	Week_Days							Total
	Mon	Tue	Wed	Thurs	Fri	Sat	Sun	
Bank	3	2	2	0	3	3	1	14
Bar	17	16	16	28	53	46	45	221
Education	29	27	37	31	26	9	10	169
Entertainment	18	17	13	21	27	36	36	168
Food	132	133	149	153	160	193	172	1092
Grocery	13	7	14	14	11	13	35	107
Gym	6	12	13	8	10	4	7	60
Home	13	12	8	11	12	9	12	77
Hotel	8	6	10	8	11	4	11	58
Landmark	9	5	15	11	9	8	12	69
Medical	8	8	10	6	11	5	4	52
Neighbourhood	1	5	4	8	14	7	13	52
Outdoor	17	3	5	8	7	21	29	90
Public Transport	29	15	28	37	22	17	35	183
Religious Venue	1	2	0	0	1	0	9	13
Salon	0	1	3	6	3	3	2	18
Shopping Centre	28	32	20	37	33	35	43	228
Sports	18	23	17	36	18	47	35	194
Work	29	26	31	37	38	25	14	200
Total	414	377	444	497	507	518	573	3330

The findings also show that as the week passes, people indulge more in leisure activities. Table 4.6 shows that check-ins are highest at venues such as Bar, Entertainment, Food, Outdoors, Shopping centers and Sports centers during weekends. In contrast, activities at work are slowest on the weekends. Airport activities are similar throughout the week with Wednesdays and Sundays a little busies than other days. If we look at the Education category, which includes colleges, universities, schools and libraries, there is not much activity during the weekend compared to weekdays. The findings are in line with previous literature that most students like to work hard during the weekdays and party hard during the weekends. Bars and pubs start getting busier from Fridays onwards, with Friday being the busiest day. A lot of people like to end their stressful week with a drink or catch up with family and friends. Sunday is the busiest day for grocery shopping with all other six days showing similar sort of activity.

One would imagine that Saturday should be as busy as Sunday because most people have days off on both these days; However, that does not seem to be the case. This can be explained with the results from the table showing activities associated with leisure such as sports, going out for meals, entertainment and bars have highest check-ins on Saturdays compared to Sundays. This shows that most people like to enjoy, relax and have fun on Saturday and leave the grocery shopping, which is more of a mandatory task than leisure activity, for Sundays. Apart from grocery shopping, Sunday is also a preferable day to meet other

shopping needs as the day has the highest check-ins for shopping centers. Similar to shopping, Sundays also seems to be a popular day for visiting religious venues such as a church.

4.4.3.2 Gender-based Activity Analysis

Based on the results shown in Figure 4.5, males represent 58% of all check-ins compared to females, who represent 40.7% of all the check-ins. These results suggest that males are 50% more likely to go out and check-in at one of the 20 venue categories.

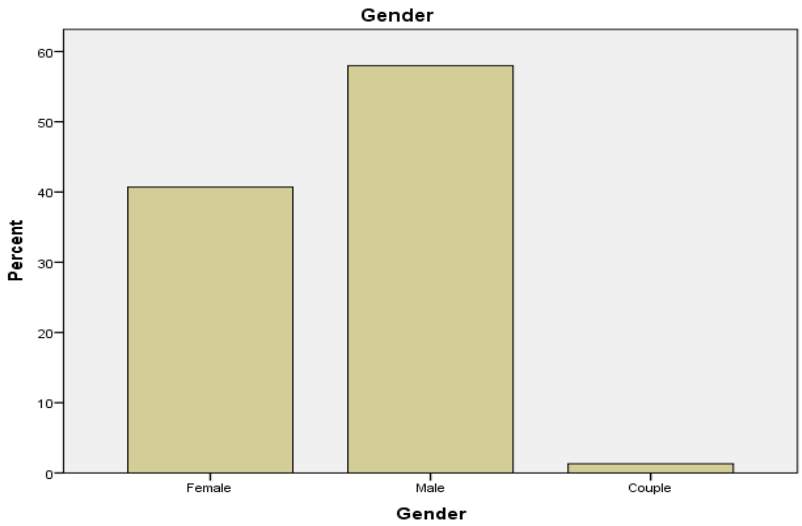

Figure 4.5 Gender-based activity analysis

Figure 4.6 shows the gender-based activity for all the 20 venue categories used in this chapter. Males lead females in 16 out of the 20 categories when it comes to visiting a place and logging a check-in. The figure shows an interesting finding; 4 out of 5 check-ins made in bars are by males. However, 4 in 5 check-ins made in shopping centres are by females. One would imagine that even men need to go to shopping centres to buy clothes, shoes etc., so why is their check-

in percentage so low at shopping centres compared to women's? Similarly, one can argue that the percentage of women going to a bar or a pub for a drink has to be more than 20% however, that is not what the results reveal.

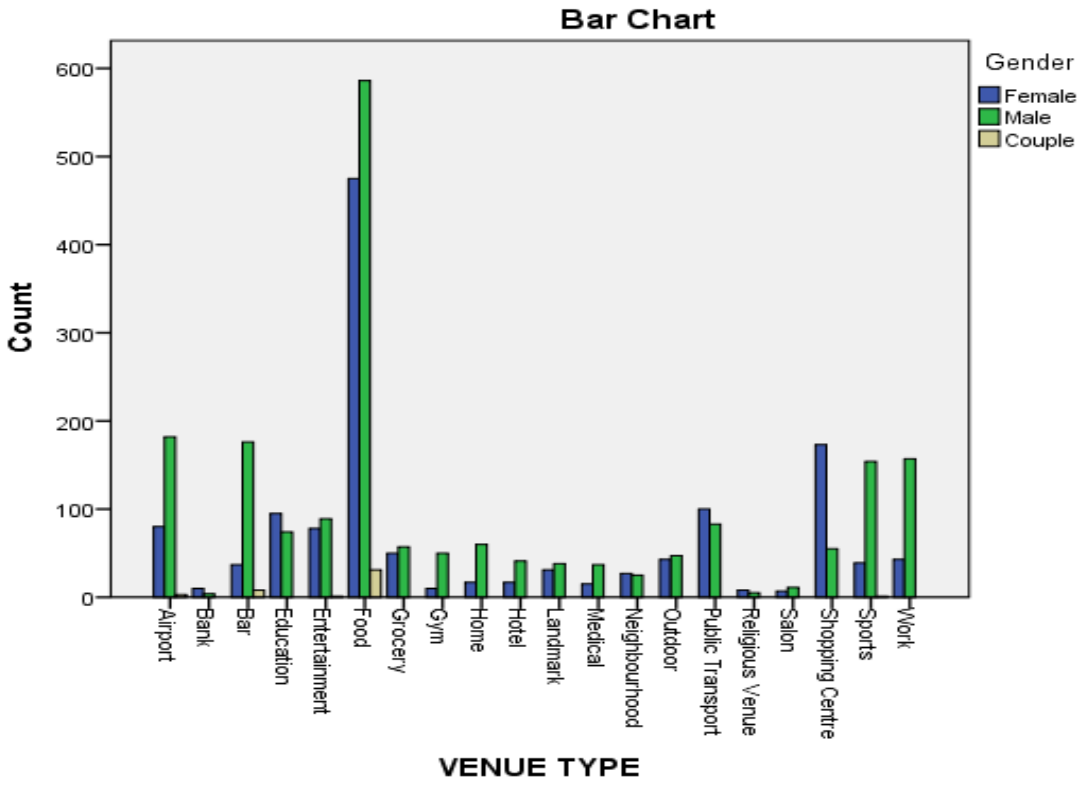

Figure 4.6 Gender and venue-based activity analysis

The explanation for this could be that due to our societal norms. Women rather than men, have been associated with household chores that include grocery and other types of shopping. The low percentage of check-ins at shopping centres by men may be due to the fact that they do not want to be associated with these chores. They may not feel proud to broadcast when they are actually involved in such activities. They may feel prouder to be associated with a drink and may

like to broadcast their check-in every time they are out having few shots with friends in bars or pubs because it is socially acceptable for men to have drinks.

Due to similar societal norms, most women may prefer not be associated with alcohol. This may also be due to family pressure and other religious norms. So, even though the percentage of women going out for drinks is most likely higher than 20%, most of these women do not like to let everyone around them to know that they are drinking. Similarly, the percentage of males going out for shopping might be more than 20%, however most would prefer not to broadcast such an association.

The Figure 4.7 shows gender-based activity during different days of the week. It can be seen that male activity is in line with the overall activity in the city, i.e. it starts slow at the beginning of a week and picks up as the week proceeds. Saturday and Sunday are the two busiest days during the week. However, for females, activities do not increase as much as the week passes and remains very similar throughout, increasing slightly on Sundays. This difference in patterns can be explained by the report published by the Australian Government's Workplace Gender Equality Agency in 2016 [194]. According to the report, women constitute only 36.7% of all full-time employees as compared to the 63.3% of male employees in Australia. The number of male full-time employees is almost double of the number of female employees. Since most people in Australia work from Monday to Friday, most men are at work during those days

and when a working week finishes on Friday, their mobility activities increase. As for women, since a large number of them either do not work or work only part-time or casually, they are more flexible during weekdays, and able to go out and visit places.

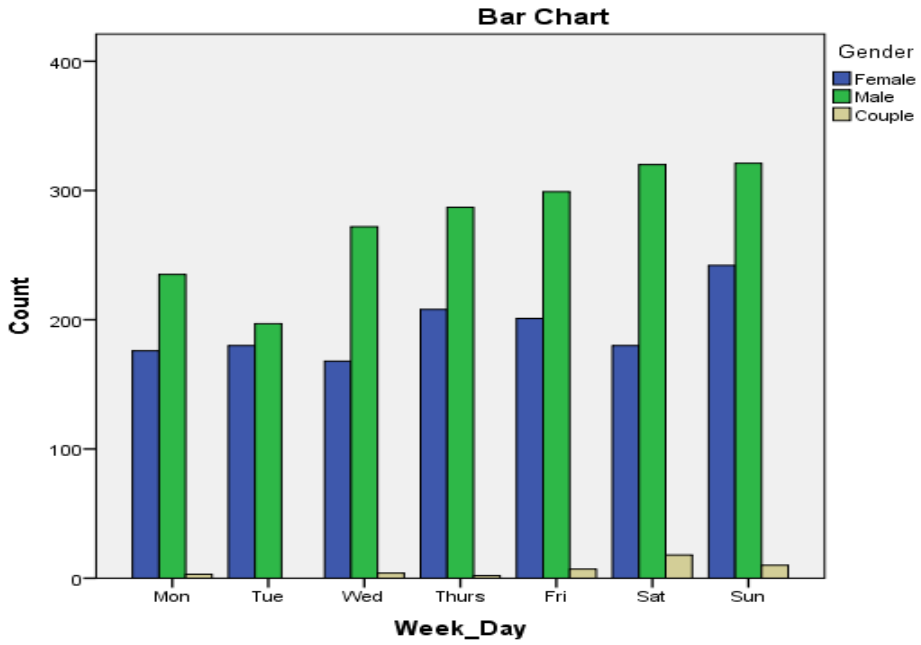

Figure 4.7 Gender and weekday-based activity analysis

The Figure 4.7 also shows that female activity is high on Thursdays and Fridays. Since four in five shopping centre check-ins are by females and shopping centres are open until late on these two days, it may imply that more women check-in at shopping centres during Thursdays and Fridays. Sunday has the highest activity when it comes to grocery shopping and it is also the highest check-ins day by females. This may indicate that most grocery check-ins on Sundays are logged in by female customers implying that females do most of the grocery shopping in households. Figure 4.8 shows the distribution of activity by male and female users in the city, based on the different time slots in a day.

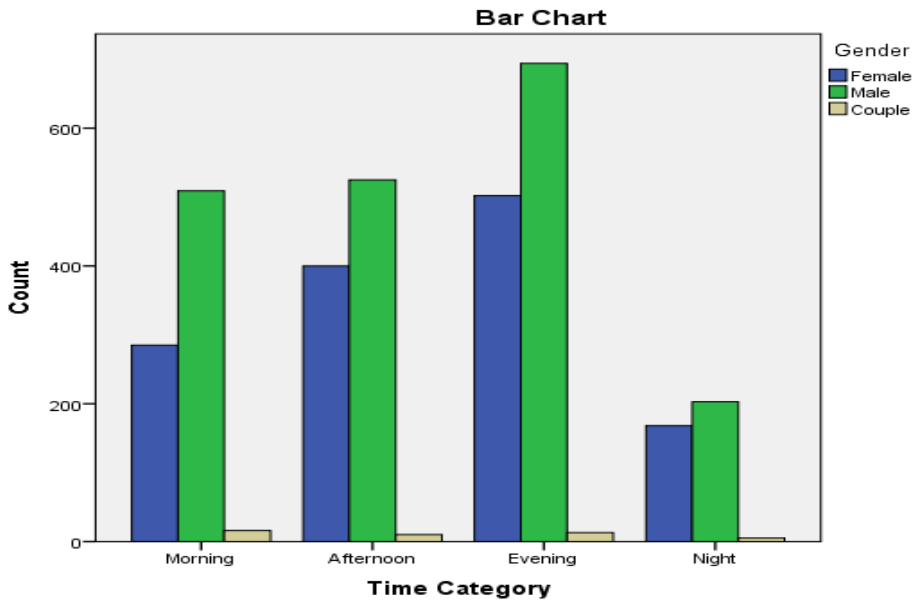

Figure 4.8 Gender and time of the day-based activity analysis

The findings reinforce what has discussed so far about the male and the female activity patterns. From the figure, it can be seen that male activities are low and very similar during the mornings and the afternoons because that is when most people, especially males (because male constitute 63.3% of all full-time jobs), are at work. Activity increases for them as they finish work and visit places and eventually goes down for the night-period. In contrast, female activity increases throughout the day and eventually falls down at night, in line with overall activities by both genders. As discussed, fewer females participate in full-time employment and that affords them flexibility to visit the places must, even during the afternoons, when most men are at work. The findings illustrate how male and female activity varies not only during the day but also during the week.

4.4.3.3 Geo-temporal Patterns

This subsection elaborates the results from the geo-temporal analysis. Figure 4.9 shows the popular places that people visit and check-in during different times of the day.

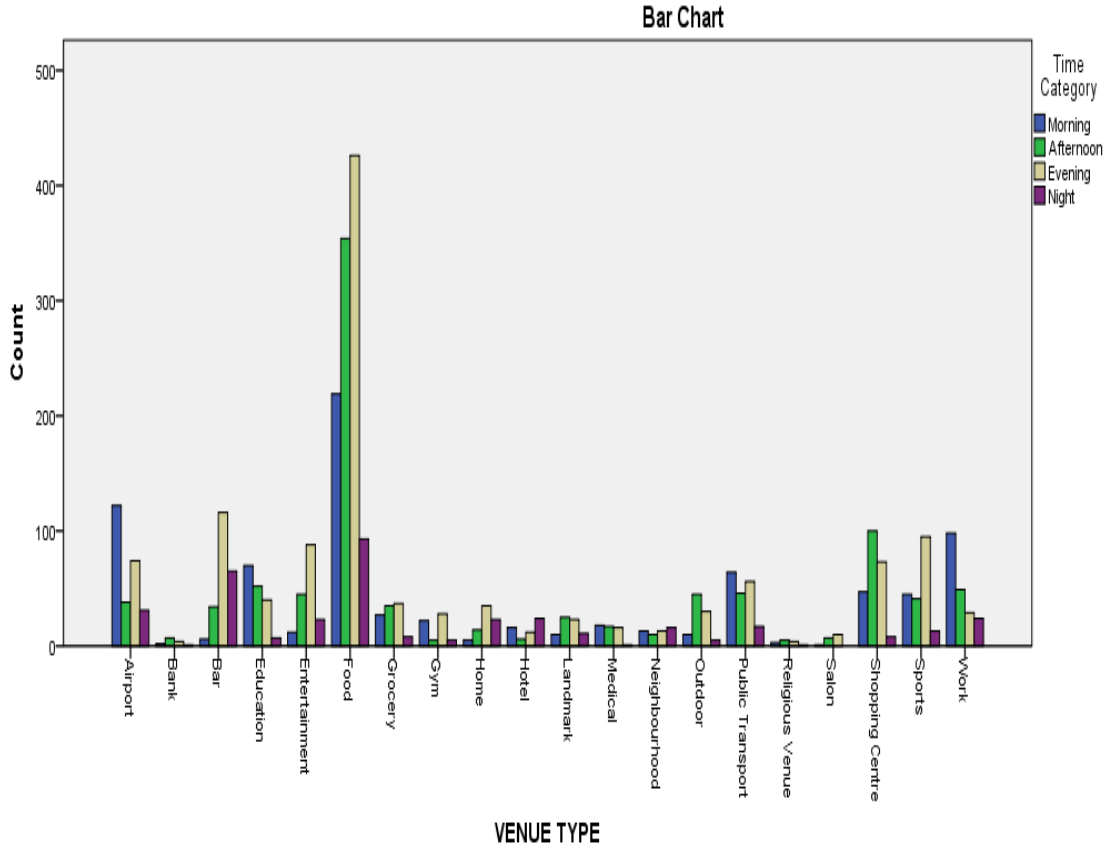

Figure 4.9 Geo-temporal pattern analysis

It can be seen that Food and Drink related places are the most popular for check-in, followed by Shopping Centres, Airport, Work, Sport Centre and Public Transport. Venues such as Religious Places, Banks and Salons are not very popular for check-ins. This may denote that some people, when visiting such venues, do not find it to be desirable to broadcast their presence there compared to being at places such as bars, shopping centres and restaurants. Surely, many

people visit religious venues and salons, however, they may not like to check-in there using Swarm. There may be many reasons for such behaviour. In regard to a bank, safety and security may be the main concern. Salons are a private place and people may not like everyone around them to know that they are having some sort of treatment to make them look better. According to Foursquare, most of its users are middle aged individuals [179] and these individuals may not find it to be desirable to be associated with religion which may explain why a lot of people do not check-in when they visit such venues.

Activity at each venue category varies depending on the time of a day. The green and mustard bars in Figure 4.9 represent afternoon and evening activities, and it can be seen that these two bars are quite prominent in almost all venue categories except in Work, Public transport, and Airport. These three categories are busy in the morning. Most people go to work in the morning and check-in when they arrive, and most Melbourne workers use public transport to get to work. This explains why morning is busy for these two types of venue categories. The higher number of check-ins at Airport indicates that most airlines organize their flights early in the morning. Whether these are incoming or outgoing flights, airport is busy in the morning. The second busiest time at the airport is in the evening. This may be due to domestic flights because most people who fly into Melbourne in the mornings, fly back in the evenings. Apart from Work, Public transport and Airport, other the categories that are also somewhat busy in the morning are Education, Food, Gym, and Hotel. Most

education institutes are busy during the morning and afternoon. Busy food places in the morning means a lot of people are getting their coffees and breakfasts. As discussed in earlier sections, mornings and evenings are the two popular time for gym goers. To conclude this subsection, different Melbourne venues are busy during different times of the day and the results shown in the Figure 4.9 presents some meaningful trends.

4.4.3.4 Check-in Dynamics

This section presents the results of check-in (mobility activity) dynamics analysis. In this section, the mobility patterns in the city are independent of venue and gender information, with Figure 4.10 showing some of the findings from the experiments.

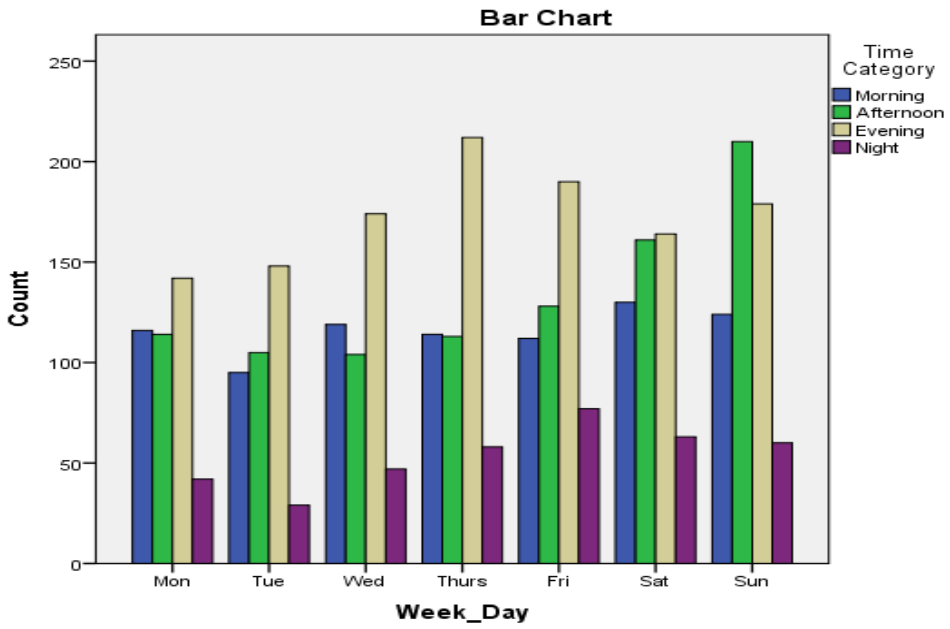

Figure 4.10 Pattern analysis

As can be seen from the results, most of the check-ins that represent mobility patterns, takes place during the evening. Evenings are busier than any other time period every day, except for Sundays. On Sunday it is afternoon activities that leads the charge. The explanation for this could be that, since Monday is the first working day after the weekend, most people tend to take it easy and stay home for Sunday evenings and in some cases prepare for work on Monday. Most people also prefer to stay home and catch up with their favourite movies on Sunday evenings. However, Sunday afternoons are the busiest, leading to relatively quieter evenings.

From the figure, it can also be deduced that night time activity is highest on Fridays, followed by Saturdays. Most people are in a 'party mood' on the weekend beginning on Fridays and leading Into Saturdays. Activity goes down on Sunday nights. The figure also shows that night time activity is slower at the start of a week and picks up as the week proceeds, with the exception of Monday nights. A lot of young people who work on the weekends, go out on Monday nights and this may explain the higher activity for the night. Morning activity is pretty stable throughout the week. So even though most people do not go to work on weekend mornings, they are still busy checking in at places such as café's, gyms, sports centre, airport, shopping centre, and outdoors. The results also show a positive correlation between afternoon activities and the day of the week. The afternoon activities rise as the week proceeds. Overall, afternoon

activity is higher during the weekend compared to weekdays as most people are at work during weekday's afternoons.

4.5 Summary and Findings

In this chapter, sentiment analysis, psychometric analysis and various descriptive statistical analyses on location sharing check-in data from Swarm was performed to explore various human mobility and behaviour patterns. The results show that restaurants, café's bars, shopping centres, gyms, and sports centres are some of the most visited places in the city, and activity at these places increases as the week passes. Sentiment and psychometric analysis of messages associated with the check-ins showed that most people do not express strong emotions related to places they visit. Descriptive analysis revealed that mobility patterns change during times of the day and the week. Night activity is slow in the city at the start of a week, picks up as the week passes, and is highest on Fridays. Activities outside work were low in the mornings and afternoons, and picked up in the evenings for male users, and remained nearly the same throughout the day for female users.

To conclude, the findings show some meaningful patterns and trends that can lay the foundation for future work. For future work in the area, deeper analyses can be performed on location-based data to discover busy parts of a city during different times of the day or week leading to the development of a predictive analytic framework on user mobility and activities. Furthermore, data from

other platforms can also be incorporated to find individual places of interests in different cities. The framework should be able to work in any city of the world. Data in the form of pictures, videos text etc. can be extracted and analyzed to deduce information. Nowadays people spend more time in front of screens than they did 20 years ago. Most of our interactions with the people around us are through screens. A lot of our daily chores such as shopping, entertainment, banking and work happen online. Since the ways we do things have changed, we need new ways to explore and research human behaviour traits. Researchers from different domains have started to explore social media as a new source of data. This methodology can also be applied to different application areas such as finance, banking, social studies, politics and healthcare. The idea behind the methodology is to use a new source of data (social media) and cutting-edge machine learning and deep learning techniques to study human behaviour in context in modern day life, which makes an extensive use of online platforms in general and social media platforms in particular.

This chapter is the first step towards developing a new methodology to extract human behaviour and personality traits from social media. The chapter contributed in establishing that online behaviour patterns of an individual are proxy for his/her real-world behaviour, and these patterns can be detected with a high accuracy. The chapter lays the groundwork for the work in the following chapters, which are focussed on detecting antisocial behaviour, a type of a personality trait, from social media. Data for the following chapters, and for the

experiments within, was collected from Twitter. Advanced machine learning and deep learning algorithms with various feature extraction techniques are experimented with to detect and analyse online antisocial behaviour.

CHAPTER 5

ANTISOCIAL BEHAVIOUR IDENTIFICATION USING TRADITIONAL MACHINE LEARNING AND DEEP LEARNING ALGORITHMS

ASB is one of the 10 personality disorders included in the 'Diagnostic and Statistical Manual of Mental Disorders [1] and falls in the same cluster as Borderline Personality Disorder, Histrionic Personality Disorder and Narcissistic Personality Disorder. It is a prevalent pattern of disregard for, and violation, of the rights of others. Online ASB is a social problem and a public health threat. An act of ASB might be fun for the perpetrator, however it can drive a victim into depression, self-confinement, low self-esteem, anxiety, anger and suicidal ideation. Online platforms such as Twitter and Reddit can sometimes become breeding grounds for such behaviour. In this chapter, a proactive approach based on NLP and deep learning is proposed to enable online platforms to actively look for signs of ASB and intervene before the behaviour gets out of control. By actively searching for such behaviour, social media sites can prevent dire situations that can lead to someone committing suicide.

5.1 Introduction

ASB is one of the 10 personality disorders included in 'Diagnostic and Statistical Manual of Mental Disorders [1]. These 10 disorders are characterized into three different clusters, and Antisocial Personality Disorder (ASPD) falls into Cluster B, along with Borderline Personality Disorder, Histrionic Personality Disorder and Narcissistic Personality Disorder [1]. It is a prevalent pattern of disregard for, and violation of the rights of others. A person with ASPD fails to conform to social norms with respect to lawful behaviour. They can become irritable and aggressive and can be consistently irresponsible when it comes to dealing with other people. The person may lack remorse and mistreat others [1, 114]. There may be many elements leading to a person developing ASB, namely genetic influences, maternal depression, parental rejection, physical neglect, poor nutrition intake, adverse socioeconomic or sociocultural factors [1, 79-83]. These factors can be categorized broadly into three main categories: Neural, Genetic and Environmental [84]. ASPD is one of the most reliably diagnosed all the personality disorders. Many psychiatrists are reluctant to treat people who suffer from ASPD because there is a widespread belief that it is untreatable however, there is increasing evidence that in certain cases it can be treated [195]. Online ASB is a widespread problem and threatens free discussion and user participation in many online communities. In many cases, it can be devastating for victims and deter them from using these platforms [196]. Online ASB appears

to be an Internet manifestation of everyday sadism. An individual who possesses and displays such behaviour online seem to feel pleasure at the expense of others, ignoring the distress and harassment their behaviour may cause [107]. Apart from sadism, attention-seeking, boredom, a desire to cause damage to the community and revenge are some of the motivations identified, and relating to the manifestation of ASB [109]. ASB annoys and interferes with a person's ability to lawfully go about his/her daily life.

Measures currently in place to discourage ASB rely mainly on users reporting it directly to platforms [120]. In most cases, victims are reluctant to confront such online behaviour as they are scared of retaliation. Therefore, most cases of ASB go unnoticed. Online platforms encourage freedom of speech but fail to draw a line between free speech and unacceptable behaviour. Current measures do not seem to prevent people from explicitly displaying ASB, exposing a lot of people (who may fall into a vulnerable group) to such behaviour.

Twitter is one of the most popular social media platforms that encourages people to share views and content. A user contributes in the form of a tweet which is a 280 word piece of text, and may contain an image, video, link to an article, etc. The platform encourages user participation in the form of discussions on topics of interest, however this may bring along some undesirable behaviour such as bullying, abuse and harassment [114]. Online ASB is prevalent mainly among users aged 18 to 27 and has also been linked to excessive use of online

platforms. The perpetrator seems to enjoy their ASB at the expense of others [107-109]. Ramifications of excessive use also lead to other psychological disorders, and employing measures to curtail their impact on society is imperative [197].

Twitter and other platforms rely on users to report ASB. Based on user initiative only, Twitter investigates the reported behaviour. The platform has automated systems in place that may prevent the distribution of illegal material, spam, nudity, pornography, etc., however, nothing is done for ASB [113]. So, on the one hand, the platform connects users to facilitate the exchange of information, ideas and other useful resources, and on the other, it facilitates the spread of ASB and related problems; putting a large number of people at risk [198].

This chapter proposes an approach based on NLP and deep learning, that can be used at a scale to classify tweets containing elements of ASB with high accuracy and precision. The approach can be implemented to automate the detection of ASB on Twitter and other online platforms, thus curbing its prevalence. To date, no prior work has either focused on critical tweet identification or evaluated deep learning and machine learning techniques against different feature extraction approaches to identify psychological disorders in general, and ASB in particular, from social media data. First, a benchmark dataset of tweets with labels, 'ASB' and 'non-ASB' was constructed. Textual features were then extracted from the unstructured data for deep

learning using cutting-edge machine learning architecture along with traditional machine learning algorithms to construct a prediction model for automatic identification of ASB. The problem of recognizing online ASB was treated as a binary text classification task in which a tweet is classified as either an "ASB" or "non-ASB" tweet, based on its textual content. Table 5.1 shows some of the examples of antisocial and non-antisocial tweets classified by using the proposed approach.

Table 5.1 Examples of antisocial tweets with corresponding labels

ID	Online Tweet	Context	Label
T_1	Fuck the law. Laws are merely words on paper backed by the promise of violence if they are not obeyed. Words on paper have no authority over any of us. Just like imaginary lines have no authority.	Failure to conform to social norms.	Antisocial
T_2	i will Personally fight the sun. come at me you fucking bitch i will Destroy You	Irritability and aggressiveness	Antisocial
T_3	Drink driving is fun and I do not mind doing it every now and then as long as I am not caught. I mean to have fun is our right.	Reckless disregard for safety of self or others	Antisocial
T_4	I snubbed the dude and then he fell. he was a looser and I could not care less. No regret	Lack of remorse as indicated by being indifferent to or rationalizing having hurt and mistreated	Antisocial
T_5	I would like to win the mega lotto this Tuesday, and if I do win a lot of people will be very happy because I am a very generous person and I like to help others good luck to me;!!!	Seeking wishes and good luck	Non-Antisocial
T_6	As a cancer survivor I like to help out where I can.	Being nice and offering help.	Non-Antisocial
T_7	So sorry for your loss. I pray for you and your family	Condolences	Non-Antisocial

The performance of the proposed approach was evaluated by comparing the results obtained from the deep learning methods and the traditional machine learning techniques. Analyses of features helped in identifying important words that could distinguish between ASB and non-ASB tweets. The experiment's results and analysis are beneficial to researchers who are interested in conducting further research into online ASB utilizing social media data.

The main contributions of this chapter are: (1) A medium scaled benchmark ASB tweet dataset with labels for ASB, and non-ASB, (2) Development of a deep learning classification model after evaluating the performance of different DL architectures, (3) Empirical validation of the higher performance of the deep learning models used in this chapter against the selected traditional machine learning models, (4) Visually enhanced interpretations of the different feature vectors in machine learning and (5) Proposition of a novel approach to study behaviour/psychological disorders from social media data using artificial intelligence.

The rest of the chapter is organized as follows. Section 5.2 provides the background on ASB disorder, aetiology, manifestation and its repercussions. Section 5.3 presents an approach to ASB tweet identification. Section 5.4 offers details on experiments conducted to evaluate the proposed approach with analyses of the results and discussion. Section 5.5 concludes the chapter and envisages future research directions.

5.2 Background

5.2.1 Online Antisocial Behaviour

To understand online ASB, it is imperative to go through the diagnostic criteria explained in the DSM-5 [1]. The term ASPD is primarily used in a clinical setting and may be used to explain the behaviour of a person who is behaving against societal norms. To be antisocial may mean to be against rules, laws, norms and other acceptable behaviour [114]. Furthermore, "against the rules and law" may refer to a failure to obey laws and the legal system, engaging in criminal activities, being arrest, etc. A person with an antisocial personality may also lie, deceive people and manipulate others for self-amusement and profit. They may easily become irritable and aggressive and be inclined to engage in fights, and they may be impulsive, irresponsible and lack remorse for their actions [84]. Not all psychological disorders can be diagnosed by a person's writing, but a few can, and ASB is one of these. Since it can be diagnosed by the way a person writes, we are able to detect such behaviour from tweets, online posts, reviews and comments. In any text, ASB is expressed by using words and their context. There are a number of rude and taboo words and short phrases that can be associated with ASB. It may seem easy for a human to pick up such behaviour through text, however, it may not be so easy for a machine [199].

One reason for this machine difficulty is that some rude words can be used in humour or in sarcasm, which may not always be considered antisocial. Also, the context of the text plays an important role in classifying it as an antisocial text.

Use of slang, the order of words, local culture, etc., all play an important role in classifying a piece of text. Some words and phrases that are normal to use in one country may imply rudeness or ASB in other. An example of this is an experience shared by a friend from Australia, who was in a café in the US and asked for a 'white coffee'. This is the normal way to order a coffee with milk in Australia however, in the US the barrister, who was a person of colour, thought that my friend was rude and racist. My friend should have asked for 'coffee with milk' instead of 'white coffee'. Under certain circumstances, it is difficult even for humans to know the exact intentions of a person from his or her writing, so we can imagine how hard this could be for a machine.

A machine or a computer relies on a set of rules and instructions to take any action however, in the case of NLP, it is not so straight forward. A few different techniques are used in NLP and, in this research project, the machine learning approach has been implemented. Training a machine learning model to detect ASB from a person's writing requires a lot of training and testing with data, along with ground truth validation. For the purpose of labelling the data set and ground truth validation, help from a psychology graduate was sought.

5.2.2 Repercussions of ASB

ASB and its impacts are both mature and well-researched areas however, online ASB is a relatively new research area and has recently gained much attention. Perpetrators often display such behaviour via cyberbullying and trolling. Targets of ASB are impacted in many different ways, and negative health impact

is one of these. Victims can suffer internalizing problems such as depression, low self-esteem, anxiety, suicidal ideation and anger [200]. They can also experience externalizing problems such as alcohol abuse, smoking, self-harm, aggression and negative behaviour towards the external environment [201, 202].

Victims of all ages can experience negative mental health impacts of ASB, however individuals who come across such behaviour as children, have higher chances of developing psychological disorders and social problems associated with it [203-205]. Some studies have linked exposure to ASB as a child to a decline in academic performance as a young child, and poor family and social relationships as an adult [206, 207]. Victims are often pre-occupied with their ASB experiences and find it hard to concentrate on academic tasks, leading to poor performance. Falls in school grades and school attendance have also been linked to exposure to such behaviour, leading to a vicious cycle affecting all aspects of academic life.

Adult victims of ASB report higher levels of anxiety, depression and severe social difficulties. The thought of experiencing such behaviour on a repeated basis prevents a lot of victims from going out and socializing. This leads to self-confinement and isolation that leads further to depression and social problems [208]. ASB at work leads to lower employee morale and lost output. Perpetrators often target their victim via inflammatory emails, offensive text messages and by posting inappropriate comments and images. Females, minorities and new

employees are often easy targets for such behaviour. Putting measures in place to manage such workplace activities and behaviour costs organizations a lot of resources, in addition to negative media coverage and higher staff turnover which add up to the cost of doing business [209].

Many studies have linked victims of ASB to drug and alcohol use, hyperactivity and a decline in pro-social behaviour [210, 211]. Victims fall prey to drug and alcohol as a convenient escape from their problems. Excessive use of substances makes victims hyperactive and discourages them from socializing. Studies have also linked suicidal thoughts and self-harm behaviour as ramifications of experiencing ASB. Self-harm may include things such as cutting oneself, jumping from heights, self-battery, burning, and poisoning; with some industrialized and developed nations experiencing higher than average incidents [212-214]. In contrast, some victims may use aggression as a way to get their frustration out and may bully, harass or troll other individuals around them [215].

Despite its relatively brief history, online ASB has been identified as a serious public health threat. Apart from the direct impact on victims and an indirect impact on their families & friends, ASB is also a burden on the public health system. The cost of treating individuals going through depression, anxiety and other related psychological disorders adds up, and impacts public health spending significantly [209, 216].

5.2.3 Obligation to Restraint

Online ASB needs to be deterred and confined. It may not be possible to eliminate it completely, however, by placing appropriate measures in place, it can be confined to a certain extent, and its impact on victims and their families can be mitigated; taking the pressure off the workplace and the public healthcare system. ASB is a huge cost to society and, with the advent of the Internet, it has become ever easier for a lot of people to indulge in such unacceptable behaviour online. The spike in incidence of ASB in general, and online ASB in particular, can be explained by the fact that perpetrators can stay anonymous online, which is not usually an option for them in a face-to-face situation [57]. In the real-world, perpetrators usually have power over victims that they exploit to bully and harass. This power can come in many forms such as social status, physical strength and workplace seniority. However, in an online world, these powers may be insignificant. Since a perpetrator can stay anonymous, he can also bully and harass someone who is higher in a workplace hierarchy, social status and physical strength [57, 63]. Furthermore, in a face-to-face confrontation, a perpetrator can cease abusing and bullying once he recognizes that he has hurt his victim enough. In online bullying, a perpetrator may not know when to stop and can push the victims toward taking extreme actions such as self-harm or committing suicide. Pushing someone online to an extent that may lead the victim to commit suicide is a form of suicide baiting, where an offender encourages a victim to take his/her own life [64, 65].

Online ASB is a huge cost and burden to our society. Damages caused by online ASB can be seen in families, workplaces and the public healthcare system. This kind of behaviour is never acceptable, whether it is online or offline. Even though large social media platforms and other online platforms have responsibilities to make sure that their platforms do not become breeding grounds for antisocial semantics, it is, however, the responsibility of all users to discourage and report such behaviour. The approach proposed in this chapter can assist online platforms to automatically identify such behaviour on a scale, and prevent it from spreading.

5.2.4 Natural Language Processing

NLP is a field concerned with the ability of a computer to understand, analyze, manipulate, and potentially generate human language. By human language, we are simply referring to any language used for everyday communication. This can be English, Spanish, French or Mandarin. A programming language such as Python, that has been used in this research, does not naturally know what any given word means. All it sees is a string of characters. For example, it has no idea what antisocial actually means. It sees that a word is a ten-characters long, but the individual character doesn't mean anything to Python, and the collection of those characters does not mean anything either. Humans know what an 'A' and a 'S' means and together those 10 characters make up the world 'antisocial', and we know what that means. So NLP is the field of training the computer to understand what 'antisocial' signifies, and from there we can manipulate or

potentially generation human language. People probably experience NLP on a daily basis without even knowing.

NLP is a broad and evolving field that encompasses many topics and techniques. The core component of NLP is extracting all information from a block of text that is relevant to a computer understanding the language. There are many techniques for NLP and machine learning methods in general, and deep learning, in particular, is the most promising of all. Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed.

5.3 Methodology

This section presents the proposed approach based on NLP and machine learning that can automatically detect online ASB and can enable platforms such as Twitter to proactively prevent it from spreading by having appropriate measures in place. Most of the research conducted on ASB has been qualitative in nature, focusing mainly on deep case study analyses. Study groups are often chosen manually and are small in number. These studies are cumbersome in nature and may require a lot of resources and time. In today's world, people spend most of their time online. Access to the Internet has changed the way people live. They spend more time in front of screens today than they ever did. Most of our daily tasks such as work, social interactions, banking, shopping, and entertainment, etc. take place online. Since the way we live and do things have

changed significantly, we need new ways to explore personality and behaviour traits [198, 217]. The research for this project has been conducted by collecting data from the social media site, Twitter. Since this data is generated during our interactions with the outer world, it has a lot of information related to our human behaviour and personalities. In this research project, such information is extracted, and used to build machine learning and deep learning models that detect online ASB.

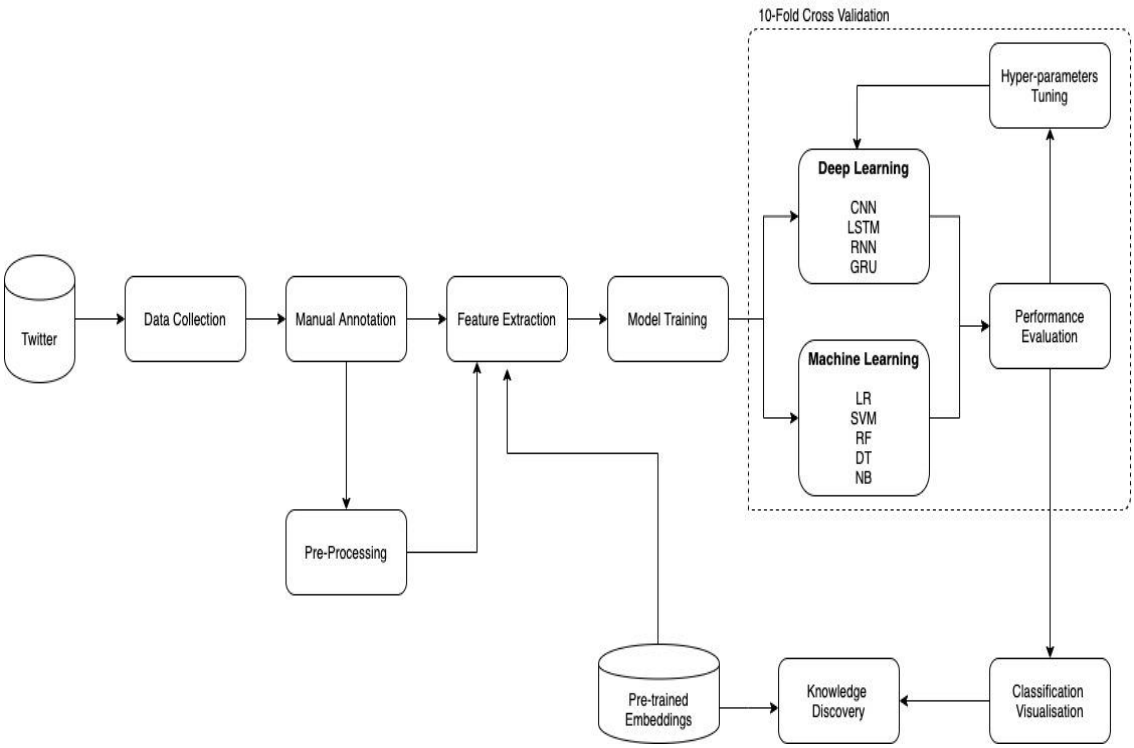

Figure 5.1 Architecture of proposed approach to detect online antisocial behaviour

The proposed approach, presented in Figure 5.1, consists of collecting the right tweets and labelling them. Once labelled, these tweets and labels must be verified by a qualified person. The qualified person in this scenario is the person who has a thorough understanding of psychological disorders and can diagnose

them in a clinical setting. Once the data (a set of tweets) was properly labelled, NLP techniques were used to clean and pre-process them. These NLP techniques are discussed in detail in the following sections. Once the data was cleaned, it was used to train and test traditional machine learning and deep learning models to establish which outputs gave the best results.

The five most widely used machine learning algorithms were experimented with to construct the final model in this chapter. These machine learning algorithms were: Logistic Regression, SVM, RF, Decision Tree, and Naïve Bayes. In regard to the deep learning models, CNN, bidirectional RNN, bidirectional LSTM, and bidirectional GRU were experimented with. Among the traditional machine learning algorithms, SVM performed the best, however all the deep learning models performed a bit better than SVM. Once the model is built it can be integrated into any online platform, including social media platforms. The model will perform well with both stored and live stream data. In the case of a live stream data, once a piece of text (tweet, post, news) is triggered to have shown antisocial semantics, it can either be removed by an algorithm or may require human intervention for further actions. A proactive approach like this can help reduce the prevalence of online ASB and encourage healthy, clean and productive online discussions.

5.3.1 Data Extraction

A total of 55,810 tweets were collected from Twitter between October 2018 and February 2019. After removing retweets and duplicates, a total of 25,500 tweets were left, and these were used in the experiments. A tweet is a 280-words piece of text that a user shares with others on the Twitter platform. When Twitter first started, the limit on the length of tweets was 140 words. It was later increased to 280 words as the popularity of the platform soared. Various phrases such as "I do not care about the law", "I wish you die soon", "Go to hell" etc., were used to search and collect tweets. Text data collected online is typically in a semi-structured or unstructured form. The data for this study was no different and was in a semi-structured form when first collected. Therefore, some tweets were missing delimiters and had no indication of any punctuation. The functions from the NLTK library of Python were used to structure the dataset. Once the dataset was in a structured form, it was annotated manually with two categories: tweets that conveyed ASB and the tweets that did not. Once the dataset was annotated, it was verified by a domain specialist from the area of Psychology. A psychology graduate was hired to do the job and the person had a thorough understanding of all the personality disorders and could diagnose them in a clinical setting.

Psychological disorders can be classified into three categories: personality disorders, behaviour disorders and state of mind disorders. Behaviour and state of mind disorders usually fluctuate and cannot be detected accurately from a

person's writing. Behaviour may change from time to time and so can state of mind. However, personality disorders or traits do not fluctuate; staying with a person for a longer period of time [1]. Since these traits stay with a person for a long period of time, these manifest through the person's speech and online writings. ASB is one such personality disorder that can be reliably detected from online corpora. The annotator was able to go through the tweets manually to see if they qualified as ASB tweets. If a tweet did, it was labelled '1'. Once all the tweets were labelled, the dataset was ready to be explored further.

5.3.2 Data Pre-processing

This phase involved removing punctuation from the tweets, followed by tokenization, which means dividing a sentence into individual words. Once tokenization was complete, stop words were removed. Stop words are the words which do not contribute much to the meaning of a sentence. Examples of such words are 'the', 'is', 'are', etc. After removing the stop words, stemming was applied to cut words into their shortest form. This is done to reduce the algorithm's computational workload. All of these steps are explained in the following paragraphs.

The first step in the pre-processing phase was to remove punctuation from the tweets. In order to remove the punctuation, Python had to be shown what punctuation looks like. This was accomplished using the String package of Python. The rationale behind removing punctuation is that the period,

parentheses and other punctuation marks look like just another character to Python, but realistically, the period does not help pull meaning from a sentence. For instance, "I like to research." with a period, is exactly the same as, "I like to research". They mean the same thing to humans, however, when these sentences are given to a machine learning algorithm, the algorithm says those are not equivalent things. A function was written to cycle through each and every character in the tweets, check if it was punctuation, and discarded it if it was. This was done to reduce the computational workload of the algorithms. By removing punctuation, the algorithms had to deal with fewer characters in the learning process.

Once the punctuation was removed from the text, the next step in the pre-processing was tokenization. Tokenization is the process of splitting a string or sentence into a list of words with white spaces and special characters. For example, take a sentence "I am doing research". Tokenization will split it into four words: 'I', 'am', 'doing', and 'Research'. Instead of seeing the whole sentence, the algorithm could see four distinct tokens, and it knew what to look at. Some of the words in a sentence are more important than others. For instance, the words 'the', 'and', 'of', and 'or', appear frequently but offered little information about the sentence itself. These are stop words. They were removed to allow the algorithms to focus more on the most pivotal words in the tweets. From the example above, if 'I' and 'am', are removed, we are left with 'doing

research'. This still gets the most important point of the sentence, but now an algorithm is looking at half the number of tokens.

The next step in the process was stemming. Stemming is the process of reducing inflected or derived words to their word stem or root. In other words, to chop off the end of a word, leaving only the base. This means taking words with various suffixes and condensing them under the same root word. For example, words such as 'connection', 'connected', and 'connective' can all be stemmed down to one base/root word 'connect'. Stemming achieves the same goal by reducing variations of the same root word and giving an algorithm fewer words to deal with. Without stemming, an algorithm will need to keep all three words: 'connection', 'connected' and 'connective' in memory, increasing the computational workload and making the machine learning models less efficient.

To summarize, the purpose of these pre-processing steps is to reduce the size of the text corpus for the machine learning model to deal with. For stemming, the Porter stemmer from the NLTK package was used. The deep learning algorithms implemented in this study do not require the same kind of pre-processing as the traditional machine learning algorithms. Duplicate tweets, websites links, and retweets still had to be removed to structure the data set, and to iron out any abnormalities when working with deep learning architectures.

5.3.3 Model Construction

The natural language toolkit is the most utilized package for handling NLP tasks in Python. Usually called NLTK for short, it is a suite of open-source tools originally created in 2001 at the University of Pennsylvania for the purpose of making NLP in Python easier. NLTK is useful as it basically provides a jumpstart to building any NLP tasks by providing basic tools that can then be chained together rather than having to build all those tools from scratch. NLTK was used for the traditional machine learning algorithms. Once the data was cleaned and pre-processed, it was converted into a form that could easily be understood by the machine learning algorithms. This process is called vectorization. Vectorization is defined as the process of encoding text as numbers to create feature vectors. A feature vector is an n-dimensional vector of numerical features that represents an object. In this research's context, it meant taking individual tweets and converting them into a numeric vector that represented those tweets.

This was done was by taking the dataset, that had one line per document, with the cell entry as the actual text message and converting it into a matrix that still had one line per document, but had every word used across all documents as the columns of the matrix. Then within each cell was counting, representing how many times that certain word appeared in that document. This is called a document-term matrix. Once the numeric representation of each tweet was obtained, the machine learning pipeline was carried out, and the machine

learning model was fitted and trained. The text was vectorized to create a matrix that had only numeric entries that the computer could understand. In this case, counting how many times each word appeared in each tweet.

A machine learning model understands these counts. If it sees a '1'/'2'/'3' in a cell, then the model can start to correlate that with whatever it is trying to predict. In this case, ASB. Algorithms analysed how frequently certain words appeared in a tweet in the context of other words to determine whether the tweet manifested ASB. In this research, both Word Frequency (WF) and Term Frequency-Inverse Document Frequency (TF-IDF) methods of vectorization were used. This was done to see the difference in the performance and the results of the machine learning model. The count vectorization created the document-term matrix and then simply counted the number of times each word appeared in that given document, or tweet in this case, and that is what was stored in the given cell. The equation for this is:

$$wf(w,d) = \frac{number\ of\ occurrences\ of\ a\ word\ in\ tweet}{total\ number\ of\ all\ words\ in\ a\ tweet} \quad (1)$$

TF-IDF created a document-term matrix, where there was still one row per tweet and the column still represented a single unique term, however, instead of the cells representing the count, the cell represented a weighting that was meant to identify how important a word was to an individual tweet. The experimentation was started with the TF term, which is the number of times a term occurred in a tweet divided by the number of all terms in that tweet. For example, if the "I like

research" sentence is taken, and the word of focus is 'research' then this term would be 1 divided by 3 or 0.33. The second part of this equation measures how frequently this word occurs across all the tweets. The number of tweets in the dataset were calculated and were divided by the number of text messages that this word appeared in, and this was proceeded by the log of that equation. For example, if there were 20 tweets and only one had the word 'research' in it, then the inverse document frequency means log (20/1).

After obtaining both parts of the equation, namely TF and IDF, the last step was to multiply both to obtain a weight for the word 'research' in the tweet. The equation is as follows:

$$W_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (2)$$

Both the matrices have the same shape, and the only difference is the values in the cells. After vectorization, the data set was ready be used by the algorithms to build the machine learning model. Machine learning is the field of study that gives computers the ability to learn without explicitly programmed. This is done by training a model using data and then testing its accuracy using more data. To this end, the dataset was divided into different buckets to train and validate the model.

In this experiment, the K-fold Cross Validation technique was used to test the accuracy of the model and tenfold cross-validation was implemented. The full

data set was divided into 10 subsets and the holdout's method was repeated 10 times. Each time, nine subsets were used to train the model and the tenth subset for testing it. The results were stored in an array and the method was repeated 10 times with different testing sets each time. In the end, the average of all test results was taken to come up with the final result. While building the model, five traditional machine learning algorithms namely Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and Naïve Bayes were experimented with. Each of these algorithms were implemented twice using two different vectorization methods: Word Frequency and TF-IDF. As discussed earlier, not much research has been conducted to deter online ASB and, therefore, it was imperative to try with both the vectorization techniques to achieve an optimum result with our dataset.

For deep learning algorithms, Keras [218], which is an open-source neural network library written in Python, was used. The four most widely used algorithms for text analysis and classification: CNN, bidirectional RNN, and bidirectional LSTM, and GRU were experimented with. Similar to the traditional algorithm, ten-fold cross-validation was used to train and evaluate the deep learning models. For feature extraction, word2vec was implemented for all four deep learning algorithms.

5.3.4 Performance Evaluation

The last stage in the model's construction was to evaluate the proposed approach to detect and identify the ASB tweets. Evaluation metrics: accuracy, precision, F-measure, and recall were used for performance evaluation. These evaluation metrics are widely used to evaluate performance for both machine learning and deep learning classifiers [40, 219, 220], and are absolute appropriate for the current problem of classifying posts and tweets related to ASB. Since one data set for the identification of ASB and non-ASB post was constructed, adopting k-fold cross validation approach was imperative, and was used. In this approach the collected dataset was arbitrarily apportioned into k partitions. Out of the k partitions, one was reserved as the test subset and the others were combined into training subsets. The whole procedure was carried out k times, which was 10 times in our scenario. The results from all these 10 folds were then averaged to indicate the overall algorithm performance.

5.4 Experiment and Analysis

5.4.1 Prediction Performance Evaluation with Traditional Machine Learning

Online ASB is relatively a new area of research. When social media platforms such as Twitter and Facebook started getting traction, they brought some issues along with them. ASB is one of them. The issue has not, so far, gained the attention it deserves, and there hasn't been much work done to detect and prevent ASB online. There are studies on cyberbullying and trolling, which can

fall under the umbrella term of anti-social, however not much has been researched on the detection of other aspects of such behaviour. By using NLP and machine learning techniques, a reasonably good job at detecting all forms of ASB has been done in this research. Table 5.2 presents the results that were obtained after experimenting with the five traditional classifiers and count vectorization. The accuracy achieved was quite high with all the classifiers used. Precision, recall, and F1 scores were also quite similar with all these algorithms.

Table 5.2 Vectorization using word frequency feature method

Classifier	Feature	Accuracy	Precision	Recall	F1 Score
Logistic Regression	WF	99.76%	99.58%	99.66%	99.62%
Support Vector Machine	WF	99.82%	99.69%	99.73%	99.71%
Random Forest	WF	98.09%	99.20%	94.71%	96.90%
Decision Tree	WF	99.71%	99.51%	99.56%	99.54%
Naïve Bayes	WF	98.84%	98.88%	97.56%	99.04%

All five algorithms detected ASB with high accuracy and precision. A tweet that was classified as containing elements of antisocial semantics was the one that contained some sort of swear or rude word designed to upset or annoy someone. Not all the tweets that were classified Positive contained swear words. The sentiment, semantic, and context of the text was also taken into consideration while manually labelling and deciding whether the tweet represented ASB.

While classifying, some of the tweets were at found to be borderline or represented more sarcasm than ASB. Such tweets were eliminated. Since this is one of the first studies trying to detect online ASB in all its different forms, borderline tweets were excluded to avoid bias and complexity. Since most of the

tweets were quite clearly Positive or Negative, the job of the classifying algorithms was a little easier as there was a limited number of words and phrases that the model had to learn to distinguish between Positive and Negative tweets.

The current study can be further extended by adding text that is more complex to classify, even by human standards. It is assumed that adding these sorts of tweets will impact the accuracy and precision metrics, however, it will enable the model to generalize better on any data set. As discussed above, the classifiers were implemented using TF-IDF vectorization as well. The results of the experiments using TF-IDF are presented in Table 5.3. SVM, when used with count vectorization, showed the best result and so did the Naïve Bayes. Logistic regression and decision tree's performance was reduced slightly when implemented with TF-IDF. Overall, the results were good, with both vectorization techniques and almost all the algorithms able to detect ASB from tweets with high accuracy. The SVM obtained the best results in all evaluation metrics.

Table 5.3 Vectorization using TF-IDF feature method

Classifier	Feature	Accuracy	Precision	Recall	F1 Score
Logistic Regression	TF-IDF	99.48%	99.64%	98.71%	99.17%
Support Vector Machine	TF-IDF	99.79%	99.77%	99.58%	99.67%
Random Forest	TF-IDF	97.76%	99.31%	94.14%	96.67%
Decision Tree	TF-IDF	99.64%	99.46%	99.40%	99.43%
Naïve Bayes	TF-IDF	93.97%	98.54%	81.55%	99.45%

Figure 5.2 presents the similarities between accuracy, precision, recall and F1 score using two different vectorization techniques: Word Frequency and TF-

IDF. In both the cases, the results are very similar. The reasons for such similar results could be the size of the dataset and the pre-processing techniques used. With regards to the size of the data set, even though the initial number of posts in the data set was around 55,000, more tweets could have bought in more variations in the data. With regards to the pre-processing techniques, Porter stemmer accomplished a good job truncating all the important words to their roots, assisting both vectorizing techniques to perform well. As can be seen from the Figure 5.2, SVM and Logistic Regression performed the best, and Naïve Bayes lagged behind in almost every metrics. Therefore, SVM was utilized for the classification model based on its performance on the dataset and its overall credibility dealing with different types of datasets.

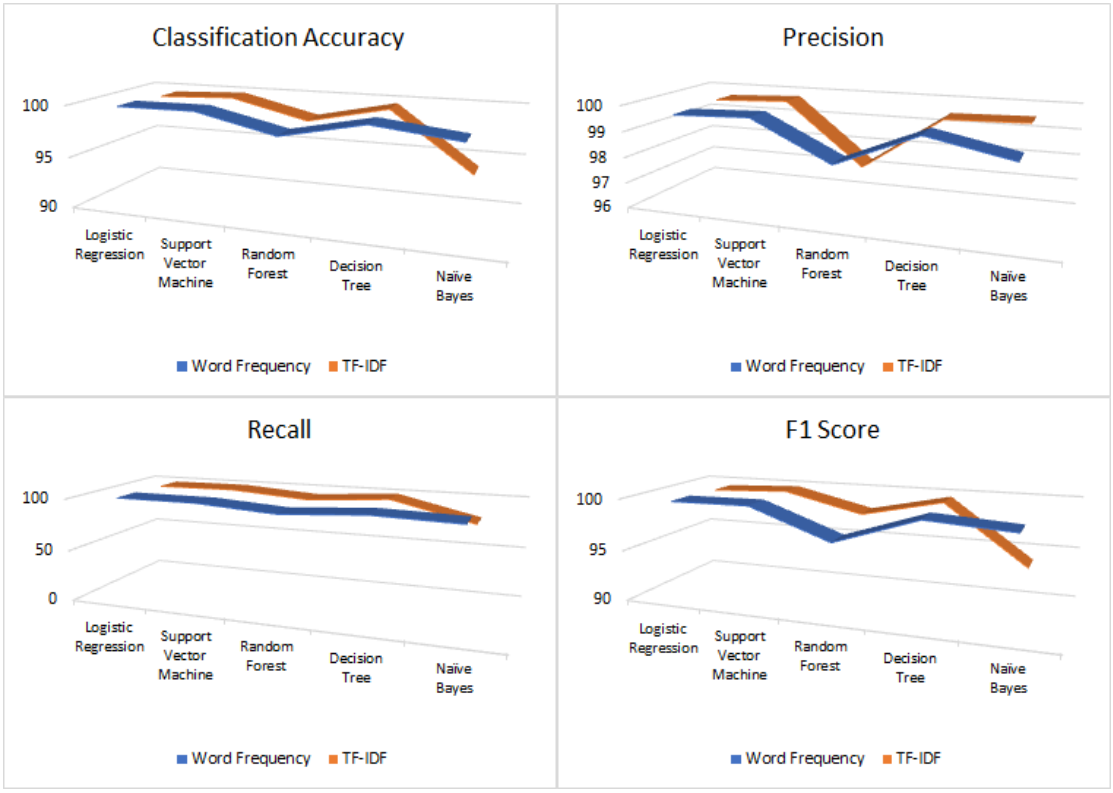

Figure 5.2 Word frequency & TF-IDF feature vector comparison

5.4.2 Performance Evaluation with Deep Learning

This section describes the four deep learning models that were used to conduct the experiment, and the results obtained by using these models. First, the inner working of these models is briefly described below.

- **CNN:** The first model used for experimentation was CNN and its detailed architecture is demonstrated in [221]. When a pre-processed tweet is fed into a CNN, it learns the embedding or the text region internally and captures the semantic coherence information of the tweet. The first layer of a CNN is known as the embedding layer and it extracts the n-gram features and stores the word embedding for each word in the text. The convolutional layer contains a disparate number of computational units and each of these units represents an n-gram from the text. Different combinations of n-grams can be experimented with, such as unigram, 2-gram, and 3-gram. The convolutional layers vary in size, and the pooling layer transmutes the previous convolutional representation to a higher abstraction level and outputs a fixed-size output. Finally, a dense layer utilizes the combination of a product feature vector to make a prediction for a tweet
- **RNN:** The next model used was RNN and its architecture is described in [222]. RNN handles a flexible-length sequence input and has loops known as the recurrent hidden state. This loop apprehends information from earlier states. At every step, it receives an input which is used to update the hidden state. One benefit of using RNN over CNN is that its hidden state integrates and utilizes information from previous time stamps

- **GRU and LSTM:** These last two models that we experimented with were bidirectional GRU [223] and bidirectional LSTM [224]. Both GRU and LSTM are improved versions of RNNs. They have memory units that maintain and store historical information, and gating units that regulate the flow of that information. There is a subtle difference in these architectures. LSTMs have three such gates whereas GRUs have two. Experiments were carried out using advanced versions of GRU and LSTM and these are called bidirectional GRU and bidirectional LSTM. Bidirectional features enable these architectures to store both future and historical information. Bidirectional features make both GRU and LSTM state-of-art semantic composition machine learning architectures for text classifications tasks.

For this study, the aforementioned four deep learning architectures were experimented with, and the results are presented in Table 5.4. 10-fold cross-validation, which was described in the previous section, was used to train and evaluate the models. Detailed performances in every iteration of the 10-fold cross-validation technique for all the four models, is presented in the table along with the averages of those iterations. The table compares the accuracy, precision, recall, and F1 scores. It also shows the epoch, which is the number of the cycle the algorithm went through to learn from the training set.

Table 5.4 Detailed deep learning classification results with epoch

Model	Fold	Epoch	Accuracy	Precision	Recall	F1 Score	Model	Fold	Epoch	Accuracy	Precision	Recall	F1 Score
CNN	1	16	0.998	0.997	0.997	0.997	LSTM	1	22	0.997	0.996	0.996	0.996
	2	9	0.999	0.999	0.999	0.999		2	7	0.998	0.998	0.998	0.998
	3	21	0.998	0.998	0.998	0.998		3	11	0.995	0.994	0.995	0.994
	4	7	0.999	0.999	0.999	0.999		4	6	0.998	0.997	0.998	0.997
	5	7	0.999	0.999	0.998	0.999		5	12	0.997	0.998	0.995	0.996
	6	11	0.999	0.999	0.999	0.999		6	9	0.997	0.998	0.996	0.997
	7	8	0.999	0.999	0.999	0.999		7	16	0.995	0.994	0.994	0.994
	8	24	0.997	0.997	0.997	0.997		8	7	0.995	0.994	0.995	0.994
	9	14	0.998	0.997	0.998	0.998		9	13	0.997	0.997	0.996	0.996
	10	14	0.999	0.999	0.998	0.999		10	8	0.998	0.997	0.998	0.998
AVERAGE		**13.1**	**0.999**	**0.998**	**0.998**	**0.998**	**AVERAGE**		**11.1**	**0.997**	**0.996**	**0.996**	**0.996**
RNN	1	9	0.996	0.996	0.994	0.995	GRU	1	36	0.996	0.996	0.994	0.995
	2	10	0.996	0.997	0.994	0.995		2	7	0.998	0.998	0.998	0.998
	3	11	0.992	0.994	0.988	0.991		3	15	0.996	0.995	0.995	0.995
	4	10	0.997	0.997	0.997	0.997		4	8	0.997	0.997	0.997	0.997
	5	11	0.995	0.994	0.995	0.994		5	6	0.997	0.998	0.995	0.996
	6	10	0.998	0.998	0.997	0.997		6	10	0.997	0.997	0.997	0.997
	7	8	0.998	0.998	0.998	0.998		7	9	0.997	0.998	0.995	0.996
	8	10	0.996	0.996	0.995	0.995		8	9	0.996	0.995	0.995	0.995
	9	10	0.995	0.995	0.994	0.994		9	9	0.997	0.996	0.997	0.996
	10	7	0.998	0.998	0.997	0.997		10	10	0.995	0.994	0.995	0.994
AVERAGE		**9.6**	**0.996**	**0.996**	**0.995**	**0.995**	**AVERAGE**		**11.9**	**0.997**	**0.996**	**0.996**	**0.996**

A lower epoch number may represent an undertrained model and a higher number usually indicates overfitting. Epoch numbers between 10-25 are considered to be a good outcome. It can be seen that the averaged epoch number for these experiments, for all the models, lies between 9.6 – 13.1. This is an indication that the models learned early on, utilizing feature vectors. Table 5.5 presents the same results in a more concise form, however, shows only the averages instead of the results from every fold, for all four models. It can be seen that accuracy and precision for all these models was close to 100%, indicating superior performance.

Table 5.5 Deep learning model evaluation

Deep Learning Model	Feature	Accuracy	Precision	Recall	F1 Score
CNN	Word2Vec	99.86%	99.84%	99.83%	99.83%
LSTM	Word2Vec	99.66%	99.62%	99.60%	99.61%
RNN	Word2Vec	99.61%	99.62%	99.48%	96.67%
GRU	Word2Vec	99.66%	99.63%	99.58%	99.60%

5.4.3 Traditional Machine Learning and Deep Learning Comparison

In this study, both traditional machine learning and deep learning models were used. Both performed well on the data set of tweets, however, deep learning outperformed traditional methods marginally. The explanation for this outperformance is that deep learning models, unlike most traditional machine learning algorithms, have capabilities to learn the semantics of a text. As explained in the earlier sections, these models have memory units and gates that can store and relay such information between different layers of architecture. These units and gates can enable complex information to be stored and communicated within the network making it possible to handle even large-scale information and assisting in learning. Learning features such as WC/TF and TI-IDF that were used in traditional methods are not capable of storing and passing information. They rely mainly on words and the number of occurrences of these words. From the traditional algorithms, SVM was the best performer and from the deep learning algorithms, CNN outperformed all the other algorithms. CNN has outperformed other deep learning algorithms in similar text classification studies in the past as well [225, 226]. So, to build a model for binary classification, based on NLP and machine learning techniques, CNN architecture is recommended to classify tweets automatically on a large scale.

5.4.4 Semantic Coherence Analysis

This subsection examines the data set to identify important words in both ASB and non-ASB tweets. ASB tweets contain mostly rude, forbidden and taboo words. They represent negative semantic and sentiments. Words such as 'F**K', 'mother**k', 'crime', 'smoke', 'lawless', 'screaming', 'bitch', 'fight', 'nigga', 'enforcement' are the most prevalent. These are not polite words and are usually avoided in social settings. One would not use such words in daily conversation unless the intention is to offend and to manifest ASB. On the other hand, non-ASB tweets are filled with positive and optimistic words such as 'respect', 'others', 'like', 'beliefs', 'help', 'grateful', 'religion', 'respected', etc. The contrast in the use of words in both classes can be seen from the word cloud in Figure 5.3. For both classes the words in large font are the ones that are prominent.

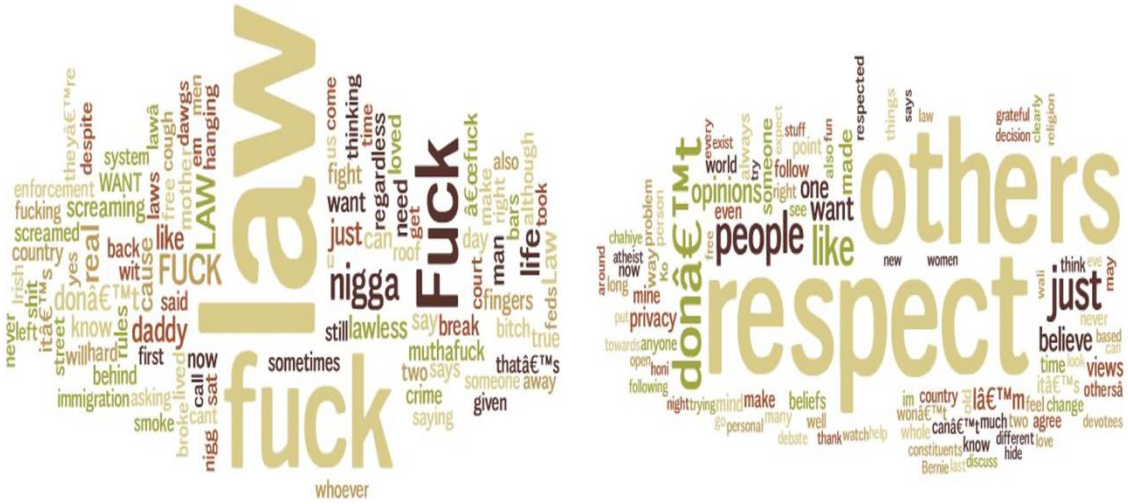

Figure 5.3 Word cloud comparison for antisocial and non-antisocial words

There is a clear distinction between the type of words associated with the ASB and non-ASB tweets. Some words appear more often in antisocial tweets than

in non-antisocial tweets and there are other words that do not appear in these antisocial tweets. Traditional machine learning algorithms rely mainly on the meaning of words and how often these words appear in a text. Both TF and TF-IDF feature extraction techniques are dependent on the meaning of words. So, even though these techniques learn to separate ASB and non-ASB tweets in the experiments, they do not fully capture the semantic relationship between these words. Deep learning, on the other hand, addresses this issue using word the embedding feature vector technique. In this technique each word is represented by a feature vector which captures the semantics of a piece of text and, because of this ability, they are able to learn better from the training data and, hence, perform better when implemented.

These word's support in identifying the underlying classes and their likelihood of occurrence in these classes was analysed in this chapter. The commonly occurring words from both the classes were taken and their supports toward identifying each tweet in that class was calculated. Some of the top words are shown in Table 5.6 along with their percentage of occurrence and Z-scores. The Z-test with p-value <=.05 was performed to establish statistical significance in difference of occurrence.

It was observed that taboo words such as shit, bitch, f**k occurred more often in antisocial tweets than in non-anti-social tweets. People who exhibit online ASB often use taboo words. Some examples of tweets are: (1) *fuck you man, I am goona*

smash your ugly face right now. (2) *I WOULD STEAL ALL OF THIS SHIT, I just cant get away wit it. Fuck the law my nigga.* (3) *Don't fuck no bitch that's fucking with your dawg, that's law. If you come up don't forget about your dawgs, that's law. I'm a street nigg@ so it's fuck the law If you broke nigg@ that should be against the lawâ*

In addition to these, words such as 'smoke', 'law', 'system', 'scared', and 'broke' are more likely not to appear in ASB. These words appear in tweets when individuals posting them claim to have broken the law, scared others, smoke marihuana, etc. Examples are: 1) *also yes . i know i shouldnt be going 60 in a 35 . fuck the law.* 2) *me & the dogs smoking nothing but nasty *cough cough*. fuck The law and whoever asking.*

Some tweets mentioned when an individual was going to break the law, hurt someone badly and threaten others. A few examples are: 1*) me & the dogs smoking nothing but nasty *cough cough*. fuck The law and whoever asking*. 2) *nah, your daddy is a real nigga, not 'cause he is hard. Not because he lived a life of crime and sat behind some bars. Cos he'll do this again for ya l!. 3) if you all gonna do what you always do, I'll be killing ya all one by one!*

The words presented in Table 5.6 exhibited some features in distinguishing antisocial and non-antisocial tweets. Solely relying on the term frequency may not be a very effective way of classifying these tweets automatically. The justification is that some taboo, bad and threating words are often used in non-antisocial tweets to either spread awareness or to report somebody to

authorities. A good classification model ought to consider the semantic relationship of words in a piece of text rather than relying solely on word count; the approach common in traditional machine learning algorithms using TF-IDF and bag of word approach.

Table 5.6 Significant difference in occurrence of prominent words

Words	Anti-Social	Non-Anti-Social	Difference	Z-Score	P-Value
Fuck	0.745	0.010	0.735	95.660	0.0000
Shit	0.450	0.020	0.430	64.048	0.0000
Bitch	0.300	0.010	0.290	50.632	0.0000
Law	0.350	0.050	0.300	47.352	0.0000
Smoke	0.310	0.030	0.280	47.079	0.0000
Kill	0.260	0.010	0.250	46.229	0.0000
Court	0.240	0.010	0.230	43.947	0.0000
System	0.280	0.030	0.250	43.630	0.0000
Hit	0.290	0.040	0.250	42.531	0.0000
Scared	0.230	0.011	0.219	42.510	0.0000
Crime	0.240	0.020	0.220	41.327	0.0000
Broke	0.240	0.020	0.220	41.327	0.0000
Enforcement	0.210	0.010	0.200	40.394	0.0000
Screaming	0.265	0.040	0.225	39.521	0.0000
Nigga	0.180	0.005	0.175	38.178	0.0000
Lawless	0.251	0.050	0.201	35.489	0.0000
Fight	0.220	0.061	0.159	28.877	0.0000
Rules	0.290	0.200	0.090	13.189	0.0000
Like	0.267	0.200	0.067	9.980	0.0000
Want	0.240	0.190	0.050	7.669	0.0000

In these experiments, the word embedding features extraction techniques of deep learning, in which each word was represented by a feature vector of 300-dimensions, were therefore implemented to overcome the shortcomings of TF and TF-IDF methods. Words with similar meaning usually have a similar

feature vector form. These 300 dimensions captured the semantics of the tweets, along with the words used in them. A word used in two different scenarios may represent a different meaning if the contexts of these scenarios are different. Furthermore, vector features of different but similar words may appear alike, and display strong correlation due the context in which they appear.

To help understand this concept better, a visualization of correlation between some of the common occurring words in the dataset is presented in Figure 5.4. The figure has the same set of words on both x-axis and y-axis. The 289 small, coloured squares represent correlations of words with other words in the figure. The diagonal from the top left to bottom right, made up of dark brown squares, shows the correlation of a word with itself. The darker the colour, the stronger the correlation, and the lighter the colour, the weaker the correlation between the words. As can be seen from the image, the word 'happy' has a high correlation with the words 'amazing', 'grateful' and 'thanks'. Similarly, the word 'asshole' has a high correlation with the words 'bitch', 'fuck' and 'shit'. The word 'behaviour' is correlated to 'attitude', and 'love' is correlated to 'happy', 'grateful', 'thanks'. The white spots show the opposite. The words 'happy', 'amazing', 'grateful', and 'thanks' have no correlation with 'bitchass' and that shows that these words fall in an opposite semantic bucket to the word 'bitchass'. Light-coloured squares show no or a weak correlation between words.

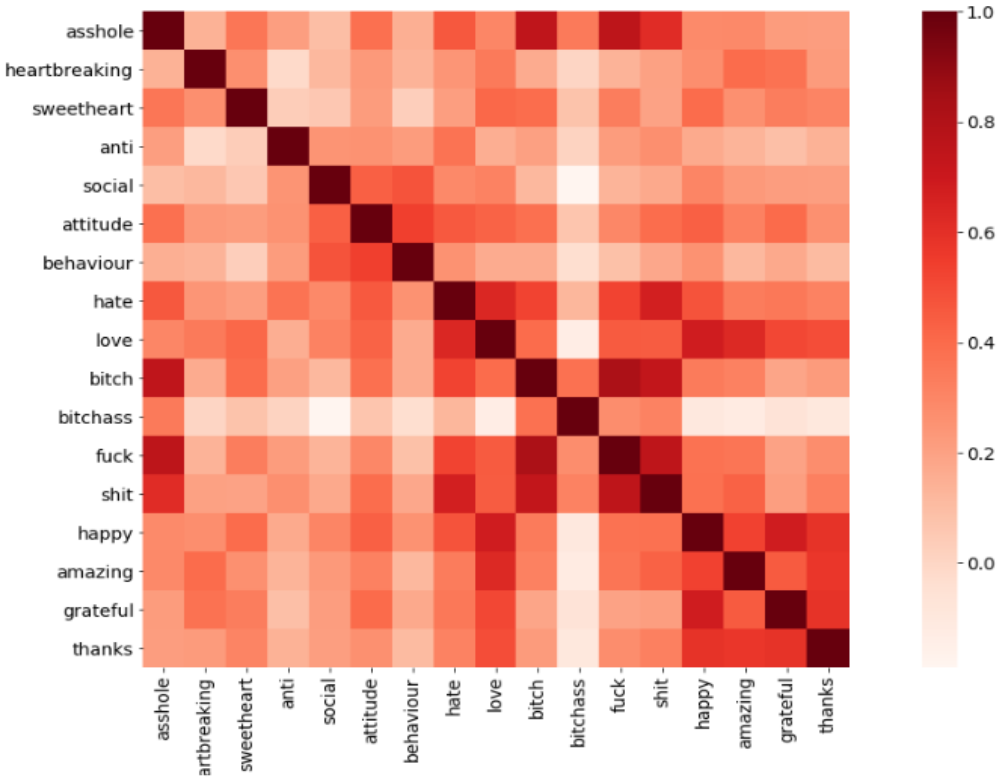

Figure 5.4 Sample word correlation

The word embedding features of deep learning models are able to capture, not just the actual meaning of words in a piece of text, but also the context in which these words are used, enabling these models to perform better when compared with traditional machine learning models. The point to note here is that, in relation to the tweet dataset, the performance of the deep learning models is slightly better than the traditional models because the models had to deal with a fewer words due to the limited size of the tweets (maximum of 280-words), however when the models are compared on larger texts such as paragraphs or even large documents, the difference between the performance widens and the deep learning models perform much better than the traditional machine learning models [225, 227].

5.5 Summary of Findings

This research introduces a data-driven approach to detect and prevent ASB online. The social media platform, Twitter, has a responsibility to prevent its platform from becoming a breeding ground for ASB. Some other online platforms also enable the spread of antisocial semantics that plague the concept of freedom of speech online. Online ASB obstructs constructive discussion and leads to many users abandoning participation. At this stage, most of these platforms rely on users reporting such ASB to these platforms instead of using an automated detection system which is imperative if prevention must happen on scale. These platforms may have some measures in place to prevent online ASB, however these measures are not as effective as they should be.

The current research proposes an approach based on NLP and deep learning techniques that can enable online platforms to proactively detect and restrict ASB. As can be seen by the results, the proposed model can detect ASB on Twitter with a very high accuracy. The model can be integrated into an online system to depict such behaviour on a live data stream. Once detected, appropriate action can be taken, such as deleting the tweet or even blocking the user, to prevent future incidents.

In this research, the data has been mainly explored from Twitter. Further studies can be conducted by collecting data from all sorts of online platforms. The diversity of data used will enable models to learn and perform better.

Furthermore, other personality disorders and behaviours that fall under the same category as ASB, can be experimented with. Diagnostic criteria for these disorders overlap in some instances and can present a challenge in training a model to classify and distinguish these disorders with high accuracy and precision.

Future work can also be carried out by classifying tweets into different types of ASBs and contexts. This may lead to offering help to victims, depending on the seriousness of the situations, by notifying authorities of the dire circumstances. Despite the findings and the achieved results, the current work has a few limitations. The size of the data set is considered moderate due to the labour-intensive job of labelling tweets manually. It consists of approximately 55,000 tweets, however a larger dataset could have brought in more diversity regarding the feature words and phrases that the algorithms used to learn from. Furthermore, around 30 different phrases were used. Once these tweets were collected, these were labelled as either antisocial or non-antisocial. The number of phrases used can be increased (to around 100) for future studies in the area. This will bring a greater diversity of words, phrases, contexts, semantics and scenarios used to train the classifier. Nevertheless, the findings and the results are valuable in guiding further ASB studies from social media data using a deep learning approach.

CHAPTER 6

DEEP LEARNING FOR MULTI-CLASS ANTISOCIAL BEHAVIOUR IDENTIFICATION

Social media has become an integral part of daily life. Not only does it enable socialization, collaboration and the flow of information, it has also become an vital tool for businesses and governments around the world. All this makes a compelling case for everyone to be on some sort of online social media platform. However, social media's benefits are overshadowed by some of its shortcomings. The manifestation of online ASB is a growing concern that hinders social media participation and cultivates numerous social problems. ASB comes in many forms such as aggression, disregard for safety, lack of remorse, unlawful behaviour, etc. This chapter introduces a deep learning-based approach to detect and classify different classes of online ASB. The automatic content classification addresses the issue of scalability, which is imperative when dealing with online platforms. A benchmark dataset was created with multi-class annotation under the supervision of domain experts. Extensive experiments were conducted with multiple deep learning algorithms and their superior results were validated against the results from the traditional machine learning algorithm. A visually enhanced interpretation of the classification process is presented for model and error analyses. Accuracy of up to 99% in class identification was achieved on the ground truth dataset for empirical validation. This study is evidence of how cutting-edge deep learning technology can be

utilized to solve the real-world problem of curtailing ASB which is a public health threat and a social problem.

6.1 Introduction

A personality disorder is an enduring pattern of inner experience and behaviour that deviates markedly from the expectations of the individual's culture, is pervasive and inflexible, has an onset in adolescence or early adulthood, is stable over time, and leads to distress or impairment [1]. There are 10 personality disorders and these are grouped into three clusters. ASB falls into the Cluster B of personality disorders along with Borderline, Histrionic, and Narcissistic Personality Disorder. Individuals who experience symptoms of these disorders often appear emotional, erratic and dramatic.

ASB is a mental health disorder that has been made popular in movies and television and there is a lot of misunderstanding and misinformation amongst the general public. There are a number of criteria that one can meet to be classified as displaying ASB. Some of the characteristics of ASB are repeated acts that violate social norms, deceitfulness and lying, impulsivity, irritability and aggressiveness, reckless disregard for the safety of self and others, consistent irresponsibility, and lack of remorse. Irresponsibility can be over one dimension, such as work and family, or across multiple dimensions. An individual with ASB, when committing an act that harms other people, does not feel guilt or exhibits any remorse. A lot of the time the person tends to blame the victim or

imply that the victim deserved to be treated that way; displaying a lack of empathy.

A number of people with ASB commit severe crimes, however that is not the only criterion for someone to exhibit ASB. Just being rude and using taboo words can sometimes qualify as ASB. ASB is extremely difficult to treat, and the charming demeanor and manipulation techniques embraced by offenders makes it even harder to deter it. There are a few behaviour characteristics are often possessed by offenders, and these are lack of empathy, superficial charm and inflated self-appraisal.

Online ASB is the manifestation of ASB on social media, blogs, news channels and various other online platforms through which participants can express their views and share information. When using these channels, individuals with an antisocial personality often display disregard for other participants and the law, use abusive and threatening language, and behave in a socially unaccepted manner. There has not been much work done on deterring such behaviour online. Some platforms intentionally let such behaviour to prevail in the name of freedom of speech however, there is a fine line between freedom of speech and unacceptable social behaviour.

To confront online ASB, it is imperative to understand its aetiology. Many factors can lead to a person developing and manifesting ASB. Some of these

factors are parental rejection, maternal depression, physical neglect, genetics and poor nutritional intake. These factors can be divided into three main categories: Environmental, Genetic and Neural. Among the environmental factors that can lead to an individual developing ASB are: exposure to violence, peer influence, family dysfunction and exposure to ASB [84]. Research has shown that living in a poor neighbourhood, being part of a disadvantaged community, not having a stable job, living in a female-headed household, and being dependent on social security are some of the other environmental factors that can trigger the onset of ASB in adults [87, 88]. A child that has been raised by biological parents suffering from ASB has a higher probability of developing ASB in adulthood [79]. Some studies have shown that if a child of parents suffering from ASB is raised by adopted parents who do not suffer from ASB, the chance of this child developing ASB is average. Although genes play a vital role in an individual developing ASB, that influence can be mitigated by changing an individual's environment [85, 86].

Neural factors related to ASB are studied through functional and structural approaches. Whereas functional studies assess the brain's core activities, the structural approach assesses the brain's morphology. Together, these studies attempt to comprehend the core neural regions that affect an individual's cognition functions including the amygdala, frontal cortex and anterior cingulate cortex [84].

In Chapter 5, which is based on the work published in [217], an approach for binary classification of online ASB was proposed. Based on content, tweets were classified as either antisocial or general/non-antisocial. This chapter goes a step further and presents a multi-class tweet categorization technique and a fine-grained insight into online ASB. The proposed approach for automatic ASB detection and classification is much more efficient than manual investigation, and can be implemented at a scale. After careful analysis, and under the supervision of a psychologist, tweets in the dataset were classified into five different classes: four classes for different types of ASB and one for general/non-antisocial category. These classes and corresponding labels are presented in Table 6.1. The categories have been identified based on the frequency of occurrence of underlying behaviours.

The main objectives of this chapter are:

- A benchmark online ASB corpora creation with multi-class annotation
- Accuracy comparison between traditional machine learning algorithms and deep learning models
- Empirical validation of the superior performance of deep learning models over traditional machine learning models
- Word2Vec embeddings versus GloVe embeddings performance analysis
- Knowledge discovery related to ASB on social media.

Section 6.2 of this chapter provides the background to online ASB, the current state of work in automatic online text classification, and the successful

application of deep learning techniques. Section 6.3 presents the methodology which covers a) data collection, b) a benchmark dataset construction, c) feature extraction from posts, d) model construction, and e) performance evaluation. Section 6.4 describes the experiment design and analyses and delves deeper into i) knowledge discovery from pre-classified classes with descriptive statistics, ii) feature extraction and model training, iii) model accuracy evaluation, iv) hyper-parameter evaluation, v) visualization of models, and vi) error analysis. Section 6.5 is the conclusion which addresses some of the limitations and proposes future directions for related research.

6.2 Background

6.2.1 Antisocial Behaviour and Social Media.

Social media communities have the potential to be supportive, punishing, or anywhere in between [228]. These communities not only offer means to work collaboratively, but also an abundance of social and entertainment opportunities. However, they can sometimes become breeding grounds for undesirable ASB. In the domain of social studies in general, and psychology in particular, ASB has been researched extensively. However, when it comes to its manifestation online via forums, social media sites, blogs, etc., the topic is still in its nascent form. It is only with the advent of Facebook in 2004, that social media became mainstream. Facebook was followed by Twitter in 2006, and Instagram in 2010. Social media has a far-reaching impact on modern-day life and has been ingrained in our work, social life, entertainment, and other crucial

aspects of our daily life. We shop, conduct business, communicate with friends and families and even entertain ourselves on social media. In a nutshell, social media has become an integral and inevitable part of modern life, and most of us use it for one reason or the other. Apart from enabling modern-day life, social media has enabled forbidden behaviours online.

Online ASB is a widespread concern that threatens user participation and free discussions in many online communities. In some cases, it can be devastating for victims and deter them from utilizing social media platforms [229]. Online ASB appears in diverse forms. Disrespect to lawful behaviour, irritability and aggressiveness, disregard for safety, and lack of remorse are some of the most prevalent forms of online ASB [228, 230, 231]. Online ASB is an Internet phenomenon of everyday malice, through which culprits seem to have fun at the expense of others' misery and distress [107]. Besides boredom, attention-seeking, malice, revenge and sadism, motivation to cause anguish is another factor leading to the manifestation of such behaviour online [109]. Imitation effect also has a role to play in the surge of such behaviour on social media. When people see other people displaying a certain behaviour trait, many will be inclined to do the same, in line with imitation theory, normalizing that behaviour trait [112, 228]. Online ASB prevents victims from going about their lawful business.

Most platforms still rely on victims to report directly to the platform for them to take appropriate action [120]. Though there is some mechanisms in place through which victims can report such behaviour to the platform, most of these cases go unnoticed as victims are often reluctant to report due to fear of retaliation by the culprit [211, 232]. Even though some victims may report such behaviour to a platform, to manually curtail ASB online is a laborious and impossible endeavour. Therefore, an automatic system that can work at a scale is required [217]. In an effort to promote free speech, most online social media platforms fail to curb online ASB. Excessive use of these online platforms has also been linked to individuals aged 18 to 27 displaying an elevated level of ASB, causing distress to others. [107-109]. Most social media platforms have some measures in place to automatically detect pornography, spam and nudity on their platform. Nonetheless, considering the devastating impact it can have on victims, ASB has not received the attention it deserves. [113]. Online platforms entice users on a promise to connect them to the rest of the world for ideas and information. However, they inadvertently facilitate the spread of ASB, putting a large number of users at risk.

To detect and classify cyberbullying automatically, using machine learning has been attempted and accomplished by numerous studies [233-237]. Similarly, online aggression has also been automatically detected [230, 238]. Trolling, the other prevailing and analogous behaviour to ASB has been automatically detected online using multiple machine learning algorithms [110, 112, 239, 240].

A numerous research studies in the recent past have employed text analysis and NLP techniques to automatically detect and classify information from social media related to domestic violence [241], emergency situation awareness [242], and trolling [239, 243, 244], etc. however, none have attempted to automatically detect and classify different forms of online ASB from social media to help prevent its proliferation. This is a multi-class text categorization problem related to online ASB that no one has undertaken yet and is addressed in the current chapter.

6.2.2 Automatic Text Classification

Online ASB detection is basically a text classification research problem that deals with processing and analyzing unstructured text data. The data could be in the form of posts, blogs, comments or tweets. NLP is a difficult task as it involves dealing with ambiguous text. The same text can have different meanings depending on the context. The whole process becomes even harder when dealing with online text that often includes misspelling, abbreviations not commonly accepted, slang, and short words. Regardless of the difficulties, researchers have applied different machine learning approaches to emotion and sentiments analysis [245], online harassment and cyberbullying prediction [235, 246, 247], crises response and emergency situation awareness [242], domestic violence crises prediction [241], etc.

Automatic text classification consists of two different procedures. The first step is feature engineering. Feature engineering is the process of extracting features from input data and its numerical vector representation. Features are the way we represent domain knowledge for the classifier. Some of the most commonly used feature engineering techniques are TF-IDF, bag-of-words (BoW) [248, 249], topic modelling features [250], psycholinguistic features [193], sentiment lexicon features [251], word n-grams[252], and word frequency[253]. The second step in text classification involves label prediction where a machine learning model is first trained on features extracted and an annotated benchmark dataset, also known as the ground truth dataset. Once trained, the model is tested on a new unseen dataset and evaluated on numerous performance metrics. This step is repeated using different machine learning models to depict the one with optimal performance. The most optimal model is then used in research and production. Some of the most widely used algorithms for text classification in machine learning are logistic regression, Naïve Bayes, SVM, decision tree, K-nearest neighbors, and RF. These algorithms suit different sorts of problem sets, and their performance relies heavily on the feature engineering process [254]. The relevance and quality of the extracted features are directly proportional to the performance of an algorithm. Occasionally, a model trained on a very precise feature extraction process fails to generalize due to overfitting, and this should be avoided at all cost [255].

Humans have an innate ability to understand words and their contexts, however that is not an ability that computers share. The widely used feature extraction techniques such as TF-IDF and BoW are sometimes not very effective when dealing with NLP due to the lack of semantic representation of text corpus and inherent over-sparsity. To overcome such shortcoming, the relatively new deep learning approach is more appropriate as it enables the capture of word meanings and their interdependencies, leading to a computer understanding of the meaning and context of a text.

Consider the examples of the following short sentences: 'Lack of remorse', 'No regret', 'absolute disregard', and 'completely indifferent'. Although these phrases are related to ASB and they all represent an analogous idea, traditional feature engineering techniques are unable to capture their semantic relationship and representation. Deep learning using feature engineering techniques such as word2vec and GloVe addresses these shortcomings. The techniques also accommodate for misspellings, synonyms, and abbreviations that are prevalent in data collected from social media, leading to significant performance improvement in machine learning text classification problems.

6.2.3 Application of Deep Learning

The progress of neural networks was stalled until recently when the deep learning technique was developed. Deep learning is a relatively new phenomenon in machine learning techniques. It [150] has shown remarkable

achievements in domains such as computer vision, pattern recognition and image processing. The rapid progression of the self-driving car industry and enterprise automation can be credited to deep learning architectures. NLP techniques and research have also been heavily influenced by deep neural networks. Applications of deep neural networks can be seen in domains such as topic classification, text classification, machine translation, part-of-speech tagging and sentence modelling.

There are two deep learning architectures: RNNs [222] and CNNs [256]. Both these architectures take the word embeddings of text data as inputs and generate feature vectors, which are numerical representation appropriate for manipulation. CNNs have been applied to question categorization and sentence-level sentiment analysis, and have shown superior performance compared to traditional machine learning algorithms such as SVMs and MaxEnts [256-258]. Likewise, RNNs have been implemented to model the text sequence in a corpus, and have demonstrated superior performance on multi-class classification [259]. RNNs are used in either their vanilla form or in one of their variants (LSTMs [260], bidirectional LSTM [224], or GRUs [223]). These variants of RNN have been used in NLP applications and have demonstrated improved performance due to their inbuilt memory architecture which stores long-range dependencies and historical information [261].

CNNs have been utilized in tweet classification problems and have outperformed the linear regression classifier with very high precision by classifying tweets into hateful and non-hateful rhetoric [262]. In another similar study [220] LSTM outperformed, not only the traditional machine learning algorithms SVM and LR, but the CNN. CNNs have been used to classify text into informative and not-so-informative in numerous research studies related to floods [40] and natural disasters [219] to assist crisis management and response efficacy. CNNs have demonstrated significantly improved performance in these scenarios compared to RF, linear regression and SVMs. In an emergency post-classification study, RNN's outperformed a SVM and CNNs in [263], LSTM outperformed a CNN [241], and GRUs outperformed both CNNs and RNNs in their vanilla form [264]. For many NLP applications such as question-answering and sentiment analysis [265], LSTMs and GRUs demonstrate better performance over CNNs [64]. In another comparable study pertinent to sentiment classification from tweets, GRUs outperformed CNNs and LSTMs [266]. Nevertheless, all the aforementioned deep learning models demonstrated superior performance and yielded better results when compared to traditional machine learning algorithms in disparate text classification problems. Though the performance of all these deep learning models is quite comparable in many of these studies, the decision to use an optimal model is dependent on the nature of an application and the manipulation of hyper-parameters.

Furthermore, deep learning has made great strides and shown promising results in various other real-time social-media applications including but not limited to crisis information detection [219], characterization of mental health conditions [267], aggressive post prediction [268], cyberbullying detection [269], abusive language pertinent to sexism and racism detection [220], fake news detection [270], clickbait detection [226], and domestic violence analogous post categorization [241]. The current research proposes the use of deep learning algorithms to detect online ASB and to conduct a fine-grained analysis of its various forms. The chapter also aims to support initiatives to curb online ASB and spread related awareness.

6.3 Methodology

The high-level methodology framework is presented in Figure 6.1 and the detailed steps are discussed in the following subsections.

6.3.1 Data Extraction

Data for this chapter was collected from Twitter using Ncaptuer, which is a browser extension. The extension works with Nvivo software whichh is commonly used in social science research for qualitative studies. The first step in data collection was to search Twitter for appropriate tweets using pre-determined phrases. Thirty-five such phrases were collected as candidate tweets. These phrases included, but were not limited to, rude, abusive and threatening words that are normally associated with online ASB. Once the

appropriate tweets were identified, they were collected using Ncapture. Ncaputer saved the tweets in a file with the extension .nvcx, which can only be opened in Nvivo. Once opened in Nvivo, the file can be exported into comma-separated value or an Excel file to be used outside Nvivo to train machine learning and deep learning algorithms. Not all tweets containing rude and threatening words exhibit ASB. Therefore, each tweet was manually selected, discarding the ones that did not fit the criteria for ASB. The DSM-5 guidelines [1] were observed during the process and it was conducted under the direct supervision of a clinical psychologist.

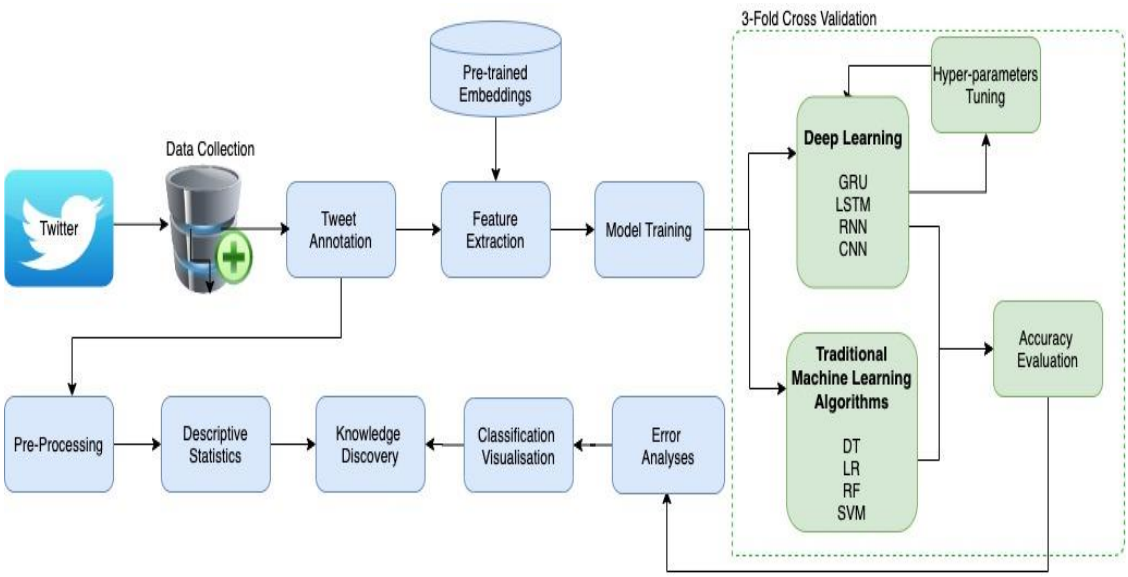

Figure 6.1 Multi-class identification proposed architecture

6.3.2 Gold Standard Construction

A gold standard dataset was constructed by manually classifying all the collected tweets. More than 25,000 tweets were initially collected and many of these were discarded as they did not meet the experiment's criteria. As abusive, threatening and rude language was searched, a many retrieved tweets exhibited only sarcasm or were written more as a joke, rather than to exhibiting ASB.

These were manually filtered out and only those tweets that clearly displayed ASB were kept. The borderline tweets were also excluded for training and testing purposes. After filtering and manual annotation, a benchmark corpus of 5,504 tweets was created. These were assigned to five classes, each representing a certain category of ASB as shown in Table 6.1. The dataset size is considered moderate, given no such study to detect and classify ASB into separate classes has been conducted.

Table 6.1 Class labels

Class	**Label**	**Number of Tweets**
Failure to conform to social norms	*1*	1192
Irritability and aggressiveness	*2*	1238
Reckless disregards for safety	*3*	804
Lack of remorse	*4*	868
Non-Antisocial or General Tweets	*5*	1402

To illustrate the annotation process further, the exemplary tweets and their corresponding class labels and classification rationale are presented as follows.

- **Class 1:** Failure to conform to social norms concerning lawful behaviour. This includes individuals breaking the law or suggesting breaking it in the future
- **Class 2:** Irritability and aggressiveness towards others displayed by some sort of aggression
- **Class 3:** Reckless disregards for the safety of self and others. Not caring for human life (personal or others)
- **Class 4:** Lack of remorse as indicated by being indifferent to others' feelings or rationalizing having hurt and/or mistreated

- **Class 5:** Non-antisocial or General class that does not exhibit any sort of ASB.

Annotation was performed (following the DSM-5 guidelines) with the help of two research students, and under the supervision of a clinical psychologist who specializes in ASB disorder. The involvement of a domain specialist was deemed paramount to ensure the accuracy and credibility of the annotation process. To substantiate the inter-rater reliability, the Kappa coefficient, which is one of the most frequently used metrics for similar studies, was calculated. The obtained degree of agreement was 0.91. In the case of a discrepancy, the class was assigned by following advice from the domain specialist. The tweets that were borderline and not fully matched by the annotators were left out of the study for consistency's sake. Some examples of classified tweets and their associated labels are presented in Table 6.2.

Table 6.2 Sample classified posts with labels

ID	**Antisocial Posts**	**Context**	**Label**
P_1	It's like they know.That I'll set fire to things, that is. I don't own any welding gear	Reckless disregards for the safety of self or others.	*3*
P_2	I feel so illiterate when responding to this idiot. The most poignant thought I can muster is Fuck You asshole.	Irritability and aggressiveness	*2*
P_3	Me and the dogs smoking nothing but NASTYYY, do I care about the law? I don’t believe sooooooooo	Failure to conform to social norms concerning lawful behaviour	*1*
P_4	I love pissing people off, and then telling them to have a wonderful day. Lmfao	Lack of remorse as indicated by being indifferent to or rationalizing having hurt and mistreated.	*4*
P_5	Today I pray for you a heart free of sadness, a mind free of worries, a life full of gladness, a body free of illness & a day full of God	General non-antisocial	*0*

6.3.3 Feature Extraction

An advantage of using deep learning for NLP applications in general, and multiclass text classification in particular, is the availability of word embeddings for feature extraction. Word embeddings are text converted into numbers that can be used by deep learning models as these models are unable to process text directly. Technically, it is the conversion of a text corpus into a feature vector that encapsulate the semantic relationship between words within the corpus; making it is an eloquent representation of text data. In word embedding, the mapping of words takes place in such a way that similar words or similar concepts appear close to each other in the vector space, disregarding any misspelling and shortcuts. The abilities of word embeddings to retain the semantic representation of text, automatic feature extraction and significant dimensionality enables deep learning models to perform better than traditional machine learning algorithms and their associated feature extraction techniques; not only by capturing semantic representation but also by overcoming sparsity. For example, in the BoW model, the terms 'remorse' and 'regret' are considered two distinctive features and will be counted separately however, in word embeddings, their position in the vector space will be very close and will bring in similar semantics to a sentence, as these both imply an analogous concept.

All most all the traditional machine learning algorithms use feature engineering techniques such as word frequency, TF-IDF, and BoW, and these classifiers usually overlook the semantic relationships with similar meaning words,

leading to inferior performance compared to word embeddings when dealing with NLP applications. Most widely used word embedding techniques are trained on very large external text corpus and have shown tremendous results in various text classification problems. These techniques are Word2Vec by Google [271], GloVE [272], by Stanford University and FastText [273] by Facebook AI Research.

In this chapter, the two most popular and widely accepted word embeddings, namely GloVe and Word2Vec have been used. Word2Vec has been trained on more than 100 billion words and these words were taken from Google News. The words were then mapped to a 300 dimension vector space to construct a vocabulary consisting of 3 million phrases and words [271]. The GloVe word embeddings were trained on more than 840 billion words and these words were extracted from Twitter posts. The words were then mapped to a 300 dimension vector space to construct a vocabulary of 2.2 million phrases and words [272]. Therefore, in both the word embedding, every word is mapped to a vector with 300 dimensions.

6.3.4 Model Development

In this chapter, four different deep learning architectures have been utilized. They are:

- **CNNs:** The detailed working of CNNs is described in [221]. The n-grams with most useful information are extracted in the first layer of this model

followed by word embedding storage for each word. These are then passed through a pooling layer that produces feature vectors. This convolutional representation is subsequently transformed into an abstract view of a higher level. Finally, the combination of the composed feature vectors is fed into the dense layer that produces a corresponding prediction for a text corpus or, In our case, a post

- **RNNs:** The detailed workings of RNNs are explained in [222]. These networks take an input of variable length sequence using a loop known as the recurrent hidden state. The loop captures information from the previous states of neurons. At every timestamp, a neuron receives an information input and updates the hidden state. Sentences are just sequences of words, and the order of these words matter to fully understand the contexts and semantics. The structure of sentences and how words are put together conveys a comprehensive understanding of semantics compared to just counting those word individually without any context. RNNs can use the ordering as a model, effectively making them well-suited for NLP problems. Therefore, the main advantage of using RNNs over CNNs is that the hidden state within them integrate information from previous time stamps
- **LSTM and GRUs:** GRUs [223], LSTMs [224, 260] are both enhanced versions of RNNs. The fundamental idea behind LSTMs is that their memory units capture and store historical information over time. Their non-linear gating units regulate the flow of information between neurons and layers. GRUs are essentially LSTMs, however, unlike LSTMs which have three gates, GRUs

have only two gates. GRUs combine the 'forget' and 'input' gates into one consolidated unit known as the 'update gate. LSTMs are able to integrate contextual information from previous timestamps, enabling the hidden state to capture and utilize this information. Due to these capabilities, both GRUs and LSTMs are regarded as cutting-edge semantic composition architectures well suited to various text classification problems. These architectures learn and capture long-term semantic and contextual dependencies between words in a text corpus and disregard any information that is redundant.

6.3.5 Performance Evaluation

Accuracy, precision, F-measure and recall are the metrics used to evaluate the performance of the classifiers used in this research. These are widely accepted evaluation metrics for machine learning algorithms [219, 220]. K-fold cross-validation was also adopted to enforce robustness to the validation process and impede selection bias and model overfitting which are common problems that can occur when trying to improve efficiency with fine-tuning features [274]. In K-fold validation, the dataset is randomly split into k number of sets. One of these sets Is used for testing and others are for training and validation. The whole process is repeated k times using different training sets, and the results are averaged to obtain the final performance metric of an algorithm or model.

6.4 Experiment Design and Analysis

This section discusses the process of automatic classification experiments for category identification from ASB posts. To evaluate the performance of the proposed deep learning approach, several steps were performed, and these include:

A. **Descriptive training:** The training procedure for both traditional machine learning and deep learning models is described, incorporating the rationale in parameter settings. Both Word2Vec and GloVe models are explored, and feature engineering techniques are presented

B. **Accuracy evaluation:** The most widely accepted validation metrics, i.e. accuracy, recall, precision, and F-measures were calculated and compared on our benchmark dataset. All four deep learning architectures, namely RNNs, LSTMs, GRUs, and CNNs, were evaluated with these five metrics. For comparison purposes, traditional machine learning algorithms such as RF, SVM, LR, and DT, were also experimented with

C. **Hyper-parameter evaluation:** Considering the impact related hyper-parameters can have on the performance of any algorithm or architecture, a number of experiments were conducted by adjusting various parameter-settings. The parameters adjusted for optimization were dropout rate, optimizer type, word embeddings, number of memory units or convolutional filters, and the number of recurrent units. Since training and

tuning an artificial neural network can be quite time consuming, the study by Reimers et al [275] was followed for the chosen parameters

D. **Model visualization**: The confusion matrices and scatter plots presented illustrate the output performance of all the deep learning architectures implemented. Graphical representation aids in understanding, not only the similarities between the different classes, but also misclassifications during the training and testing process. These visualizations present an overview of the classification outputs and assist in understanding the sources of errors. Overall, these are helpful in facilitating the interpretation of model performances

E. **Error analyses:** Examples of misclassified tweets along with word embeddings of the commonly occurring terms in the dataset are presented in this section. A thorough investigation of the correctly and wrongly classified tweets afford the opportunity for active learning and classification refinement.

6.4.1 Descriptive Statistics

Fine-grained descriptive statistical analyses were performed on the text dataset for comparative purposes, and to facilitate knowledge discovery from different classes. A number of pre-processing tasks were carried out: (i) nil pre-processing, (ii) stop word elimination and (iii) application of Stemming. The total number of words for each class were counted along with the maximum and the average number of words in tweets. Finally, the most prominent and

frequently occurring terms in each class were extracted for the purpose of having a deeper understanding of the nature of disparate classes. The results are shown in the Table 6.3.

Table 6.3 Exploratory data analysis of all classes

Pre-Processing Steps	Word Count	Failure to conform to social norms with respect to lawful behavior	Irritability and aggressiveness	Reckless disregard for safety	Lack of Remorse	Non-Antisocial / General
No Pre-Processing	Total Words	21259	23912	23656	22718	33115
	Max. Word Count of Tweets	63	65	67	93	61
	Average Word Count of Tweets	18	19	29	26	24
	Most Common Words	he,fuck,system,law,i,and,to,a,you,it,my,this,of,is,that,they,in,for,...,all	you,fuck,i,fucking,the,to,and,a,asshole,bitch,my,of,me,fuckin,it,your,that,is,in,this	i,and,to,the,my,not,a,it,do,on,in,have,me,is,of,safety,with,am,for,that	you,i,and,to,die,have,hope,your,a,that,the,it,in,not,am,do,are,of,for,no	you,i,luck,wish,to,and,the,for,a,in,your,of,it,that,but,with,is,on,my,this
Without Stop Words	Total Words	12379	13064	10523	9510	17138
	Max. Word Count of Tweets	54	41	30	33	53
	Average Word Count of Tweets	10	11	13	11	12
	Most Common Words	fuck,system,law,...,da,em,make,free,like,get,shit,people,i'm,life,nigga,pay,want,fucking,earn	fuck,fucking,asshole,bitch,fuckin,cunt,ass,like,whore,shit,get,mother,motherfucker,...,fucker,say,go,bastard,people,..	safety,get,almost,like,fun,killed,one,care,fast,life,fire,dangerous,hit,got,know,people,driving,car,head,set	die,hope,love,people,pain,pissing,suffering,deserve,suffer,happy,care,regret,life,know,like,hurt,death,remorse,time,see	luck,wish,pray,like,...,good,respect,others,know,help,love,great,people,get,i'm,god,one,well,hope
Stemming Applied	Total Words	21259	23912	23656	22718	33115
	Max. Word Count of Tweets	63	65	67	93	61
	Average Word Count of Tweets	18	19	29	26	24
	Most Common Words	the,fuck,system,law,i,and,to,a,you,it,my,this,that,of,is,they,in,for,...,what	you,fuck,i,the,to,and,a,bitch,asshol,my,of,me,fuckin,your,it,that,is,in,this,off	i,and,to,the,my,not,a,it,do,on,in,have,me,is,of,safeti,with,am,for,kill	you,i,and,to,die,hope,have,your,a,that,the,it,in,suffer,not,am,do,are,of,for	you,i,luck,wish,to,and,the,for,a,in,your,of,that,it,but,with,is,on,my,this

Table 6.3 shows that the total number of words declined significantly after eliminating stop words. In Classes 3 and 4 the word count was less than half, and in Class 0, almost half. This indicates that the generic vocabulary takes up a significant proportion of ASB tweets in all classes. Word count after stemming remained very similar as no words were eliminated, instead they were truncated. Furthermore, from a knowledge discovery and interpretation point of view, the application of stemming appeared futile. Due to the nature of the words used in antisocial tweets, a lot of the words did not change much with stemming. Even those that changed, such as 'safety' to 'safeti', would not have

contributed much to the performance of the algorithms because of their lack of meaning.

The notable differences between the classes were the average and total number of words in tweets. Class 4 (lack of remorse) had the maximum number of words in a tweet, whereas Class 0 (non-antisocial) and Class 1 (failure to conform to social norms) had the minimum. The average number of words in tweets falling in Classes 3 and 4 were significantly higher than the average number of words in Classes 1 and 2. This may imply that people use fewer words and shorter sentences to express irritability and aggressiveness. Some tweet examples are 'go to hell', 'I'll bash you', etc. Furthermore, it may also indicate that people write lengthy posts if they want to display their disapproval and disregard towards others and their safety. They may feel the need to justify their ASB for self-satisfaction. Some examples are:

'I love pissing people off who are jerks, some guy wouldn't wait his turn on the plane and almost knocked Cherie and raven over, so I proceeded to block the walkway with my two bags so he couldn't pass, he wasn't happy but didn't say shit, you know because I am crazy.'

'just read oomf tweeting that all antifascist people aren't even people and honestly i just hope he chokes on fish spine or maybe shoot his own head with the gun i know he has fucking bolsonaro supporter i hope you die.'

These are some important distinctions in the way people express their behaviours online, and this is often also the case in the real world, where people use fewer abusive and taboo words to unload their anger, however feel the need

to explain their feelings and disapproval when showing disregard and a lack of remorse for others. The average word count in Classes 3 and 4 is one and a half times the average word count in Classes 1 and 2. The findings such as people write lengthier posts when expressing disregard for the safety of others and to display a lack of remorse, demonstrates the need for further data mining and knowledge discovery.

Overall, with and without stemming did not show much of a difference in the word count and the type of words in tweets. Without removing stop words, the most frequently occurring words included pronouns, prepositions and articles, and these can be observed in all five classes. However, after removing stop words, interesting and valuable insights relating to each class and underlying words and phrases emerged. Another notable difference that can be observed among classes is the similarities and dissimilarities in the use of words. The findings from the most frequently used words in all the classes are presented below.

- **Failure to conform to social norms concerning lawful behaviour:** Apart from the taboo words that are prevalent in the Class 2 (Irritability and Aggressiveness) category of tweets, the most commonly occurring words are related to law and the legal system. These words are expected in this class since we are dealing with unlawful and illegal behaviour. In addition, the occurrence of the first and third person pronouns such as 'he', 'her', 'they', 'I', 'my', etc. are predominant in this class. One explanation for this could be

that people express their own grudges more towards the legal system than others. Some example are: '*because they took my freedom away*', '*it was not my fault*', and '*I was right'*, etc. This demonstrates the importance of some of the stop words in the classification process

- **Irritability and aggressiveness:** This class contains of the highest number of aggressive, abusive, taboo and angry words. This is in line with the characteristics of the class. Most prevalent are 'fuck', 'fuckin', 'asshole', 'bitch', 'whore', 'shit', 'bastard' and 'motherfucker', etc. It is widely presumed that people often use abusive words to express their aggressiveness and irritability and that is why their prevalence is highest in this category. The presence of first and third person pronouns is significantly lower compared to the Class 1 tweets.
- **Reckless disregard for safety:** The words in this class significantly deviate from the words in the previous two classes discussed. The use of abusive and taboo words is almost non-existent in this category of tweets. Instead, ASB is represented using words such as 'kill', 'fire', 'dangerous', 'hit', along with some fun words such as 'like', 'fun', 'fast', etc. People writing these types of tweets seems to have fun at the expense of their own safety and the safety of people around them. The following tweet sums up the behaviour expressed in this category: "*Woah! Dodged a bullet big time. Ran through red light with no P plate on and just got a warning letter instead of a fine. Thanks NSW Govt!*".
- **Lack of remorse:** Since this category of tweets relates to the lack of regret after having hurt or mistreated someone, it consists of terms both negative

and positive in nature. Negative words such as 'suffer', 'die', 'remorse', 'hurt', 'pissing' and 'pain' are representative of having done something wrong or mistreated others. On the other hand, words such as 'love', 'happy', 'like' may represent the display of indifference and discord after having hurt. An example of a tweet from this class is: "*I'll be laughing when you'll be dying from a curable disease*".

- **Non-antisocial/General:** The characteristics of this class are the nice, non-aggressive, non-taboo and non-abusive terms, namely: 'love', 'wish', 'help', 'pray', hope, God, others, well, respect, etc. This class is clearly distinctive from the other four due to the absence of abusive and taboo words, making it a little easier for all algorithms to identify tweets in this category with high accuracy. The class consists of tweets sharing news, greetings, discussing everyday topics and, in some instances, soliciting business opportunities.

6.4.2 Model Training

The two widely used word embeddings, namely Word2Vec and GloVe, were used to extract features for deep learning models to gauge and examine their robustness. The first layer in a deep learning model is the embedding layer. By parsing the pre-trained embeddings, this layer executes the index mapping for all the words in the vocabulary and transforms them into dense, fixed-size vectors. The successive layers consist of 128 memory cells; the number of memory cells commonly used in earlier studies [275]. The models were implemented using Keras [218], a layer built on top of the TensorFlow library

from Google [276], and were trained up to 25 epochs to achieve the highest performance.

Unlike traditional machine learning models, in deep learning models, no pre-processing of text is conducted and the whole tweets were fed into the models. In any language, stop words in the text can hold valuable information that can be leveraged to boost model performance. Also, the words in the text were not stemmed. This was avoided to preserve the semantics of each sentence in its original form, to help the model understand the context better. For example, the words 'aggression' and 'aggressive' can bring in disparate contexts to a text. Initially, the Nadam optimizer was utilized, and the batch size was confined to 32 posts, considering the moderate size of our dataset. The number of recurrent units was set to 128 and the activation function used was 'Relu'. The dropout rate was fixed to 0.2 [275]. The 'dropout' is a simple and efficient way to regularize any deep neural network and to prevent overfitting [277]

With regards to the traditional machine learning algorithms, the same Word2Vec and GloVe embeddings were adopted. To overcome the shortcomings of previously published work [217], i.e., model comparison using simple feature extraction techniques, thorough and comprehensive experimentation was conducted using advanced and widely used feature extraction and model compositions. Python's scikit-learn library with its default parameter settings was implemented for the task of evaluation.

6.4.3 Accuracy Evaluation

As a part 3-fold cross-validation approach, the complete dataset was subdivided into training and testing subsets. This approach has been adopted by numerous studies [278, 279]. The following three pre-processing scenarios for traditional machine learning algorithms were experimented with:

- Only stemming
- Only stop words removal
- Both stemming and stop words removal

The performance of traditional machine learning algorithms depends greatly on the pre-processing steps. The experiments indicated that these algorithms performed best and achieved the highest accuracy on the dataset with stemming only (without removing stop words). In the context of online ASB multiclass classification, some stop words could be useful in classes identifications. As discussed in the descriptive statistic section, first and third person pronouns were among the frequently occurring words in tweets, so it made sense to keep them in the final dataset as these were part of the context of tweets and assisted in the classification process. The results from 'stemming only' (highest performance) experiments for traditional machine learning algorithms were compared with the results obtained from the four deep learning architectures used and these are presented in Table 6.4. Along with the accuracy of these algorithms, the other evaluation metrics precision, F-measures, and recall are also presented.

Table 6.4 Classification model evaluation metrics

Model	Feature-Set	Precision	Recall	F-Score	Accuracy
CNNs	GloVe	0.98	0.98	0.98	98.07
GRUs	GloVe	0.99	0.99	0.99	**99.20**
LSTMs	GloVe	0.99	0.99	0.99	98.98
RNNs	GloVe	0.90	0.89	0.89	89.38
CNNs	Word2Vec	0.94	0.94	0.94	94.29
GRUs	Word2Vec	0.99	0.99	0.99	98.60
LSTMs	Word2Vec	0.99	0.99	0.99	98.40
RNNs	Word2Vec	0.78	0.77	0.77	76.36
RF	GloVe	0.90	0.89	0.90	90.16
DT	GloVe	0.71	0.71	0.71	71.50
LR	GloVe	0.93	0.93	0.93	93.36
SVM	GloVe	0.95	0.95	0.95	94.99
RF	Word2Vec	0.90	0.90	0.90	90.64
DT	Word2Vec	0.73	0.73	0.73	74.10
LR	Word2Vec	0.93	0.93	0.93	93.36
SVM	Word2Vec	0.96	0.96	0.96	96.62

In general, the deep learning architecture's performance was superior to the performance of traditional machine learning algorithms, as indicated by the higher evaluation metrics yield. These algorithms, when used with GloVe embeddings, produced the highest results. RNNs lagged behind in performance using both the GloVe and Word2vec embeddings when compared to the other deep learning and traditional machine learning algorithms. RNNs' inferior performance can be attributed to the difficulty of vanishing gradients [154]. As new sequences are fed into the RNNs, information from the preceding sequence diminishes in these architectures. Nonetheless, this RNN limitation is addressed by its successors, namely GRUs and LSTMs. These succeeding versions overcome such shortcomings by efficiently capturing long-term dependencies

that are imperative when working with textual data, which is sequential in nature.

Both GRUs and LSTMs performed the best when used with GloVe embeddings, and achieved an accuracy of 99.2% and 98.98% respectively. RNNs lagged behind all other three deep learning architectures and all the traditional machine learning algorithms, except for decision tree. When looking into the use of Word2Vec embeddings, GRUs and LSTMs again stood at the top with 98.6% and 98.4% accuracy respectively, and RNNs again lagged behind. Among the traditional machine learning algorithms, SVM and LR performed the best with 94.99% and 93.36% accuracy respectively, when used with GloVe embeddings. Accuracy was 96.62 and 93.36% respectively when used with Word2Vec. Decision tree's performance was inferior to all other the algorithms used in this study regardless of word embedding combinations. Overall, all algorithms performed better when used with GloVe instead of Word2Vec embeddings, indicating a superior performance capability. From these results, it can be inferred that deep learning algorithms have performed better compared to traditional machine learning algorithms. There is a higher computing cost associated with these algorithms when compared to traditional algorithms, nonetheless, this is compensated with higher performance. The traditional models are best suited for high dimensional and sparse features vectors. It can also be inferred, that these algorithms are not very well suited to dense vector representation as used in this study (300 dimensions). The deep learning models

can efficiently leverage a dense representation of word embedding to obtain higher accuracy scores as demonstrated by the results.

6.4.4 Hyper-parameter Evaluation

This study's deep learning models' performance was evaluated with regards to the number of epochs required to achieve the best results. Too few epochs can sometimes leave a model undertrained, and too many epochs can lead to over-fitting. An underfitted model does not perform well, and an overfitted one does not generalize well. Finding the right balance it is crucial for the best performance of any deep learning model. Another disadvantage of having more epochs than required is the wastage of computing resources. Training deep learning models requires a lot of time and computing power, and to use any more than required can lead to wastage of valuable resources. So, experimenting and getting the right number of epochs is paramount. Figure 6.2 presents the training comparison of algorithms using GloVe and Word2Vec embeddings.

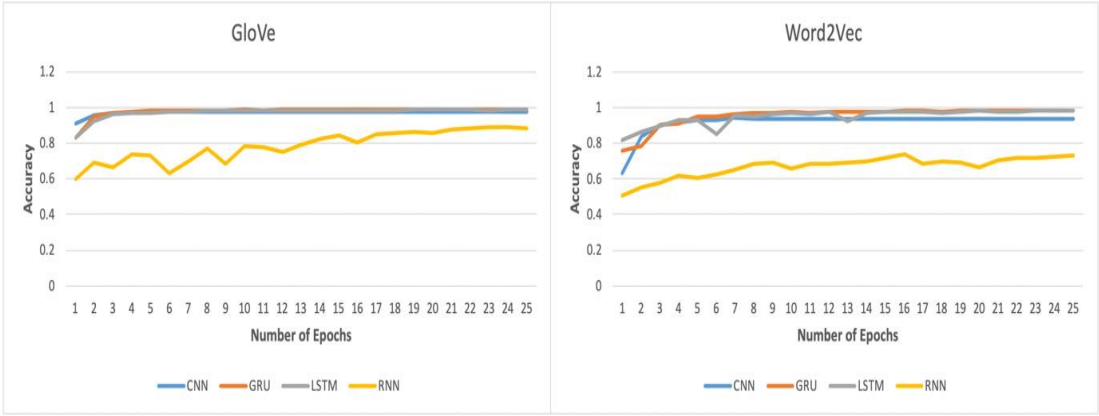

Figure 6.2 Deep learning model accuracy and number of epochs

Comparing performance when using different optimizers, Nadam and RMSProp, produced similar sort of results; with Nadam outperforming slightly because of the latter's significantly lower computation time. Performance of both SGD and Adam was inferior to the aforementioned, and SGD failed to converge in many instances due to its sensitivity related to the learning rate. In relation to batch size, it was observed that higher batch size did not relate to higher model performance. In fact, performance deteriorated with larger batch sizes. The batch size of 32 enabled algorithms to achieve better performance relative to the batch size of 256. Initially, three different activation functions were experimented with, namely sigmoid, relu and softmax. Use of these three resulted in comparable performances by the algorithms and the variance was negligible. Similar to the activation functions, the use of different recurrent units did not seem to have a significant impact on the model's performance and the setting of 128 units, which is a standard-setting, provided a slightly superior result. Considering the overall impact of these hyper-parameters on the performance of the algorithms, the following were selected and generated the best results. **Optimizer**: *Nadam*, **Activation function**: **Relu**, **Batch-size**: *32*, and the **Number of recurrent units**:*128.*

6.4.5 Models Visualization

The classification performance of deep learning architectures can be better understood with the aid of visualization. Virtualization provides insights into the inner workings of algorithms. The t-SNE dimensionality reduction

techniques, based on GloVe embedding, was used to reveal the similarities and dissimilarities among categories of ASB. Visualization indicates the categories of ASB that were correctly classified and the categories that were not. Since algorithms performed better with GloVe embedding compared to Word2Vec, visualization generated with Glove embeddings were evaluated. The highest performing models, GRU and LSTM, and the lowest-performing model, RNN, were presented for comparison. The scatter plots in Figure 6.3 show the clustering of all five classes. The more confined and distinct the clustering is, the better the algorithm has performed. The following conclusions can be drawn from the analysis of the scatter plots:

- RNN performance on the dataset was relatively inferior when compared with other algorithms used in the study. The miscalculation of a significant number of tweets can be seen. The model was unable to generate a clear distinction between some of the classes, especially between '*General*' and '*Lack of Remorse*'. The algorithm misclassified a large number of tweets that were meant to be in Class 4 (*Lack of remorse*) as non-antisocial tweets. Similarly, it also misclassified some of the Class 3 tweets (*Reckless disregard for safety*) as non-antisocial tweets. It can be inferred that the algorithm was unable to draw a clear distinction between the classes which is apparent from the lack of sufficient gaps between the clusters representing the classes
- The LSTM model performed better than RNNs and was able to draw comparatively distinct class clusters. Apart from some misclassifications, clearly defined clusters represent a decent classification performance. As can

be seen, some Class 4 tweets (*Lack of remorse*) were wrongly classified as Class 3 (*Reckless disregard for safety*). This is mainly due to the use of similar terms in both types of posts. Phrases similar to 'I don't care', 'you can die' were quite common in both classes, and led to some misclassification of posts. Similarly, Class 2 posts (*Irritability and aggressiveness*) were wrongly classified as Class 1 (*Failure to conform to social norms*). It is believed that this may, again, be due to the use of similar terms or semantics and sentiments of the posts. The algorithm was able to identify Class 1 posts with significantly high accuracy. Overall, the performance was better than RNNs

- The GRU architecture was able to distinguish posts fairly correctly, as can be seen in the scatter plot. There is a clearer distinction among classes represented by well-defined clusters with a significant gap between them. Nonetheless, there were a few misclassifications in almost all classes. Some posts from Class 1 (*Failure to conform to social norms*) were misclassified as Class 3, (*Reckless disregard for safety*). One example is:

  '*fight the powerfuck the systemkick up a mosh pit when they don t wanna listen*". Even though the post falls into Class 1, the words 'fight' may have led the algorithm to classify it as a Class 3 post. Similarly, the following post from Class 3 was misclassified as a Class 4 post, most likely due to the similarity of words in both classes, *'said this so many times and i ll say it till the day i die i do not care for my life i do not care what happens to me i care about what happens to my friends and i care about my friends lives i want everyone of them to succeed and become great'*. There were misclassifications with other four Classes as well,

however, the performance of GRUs was significantly superior to the other algorithms.

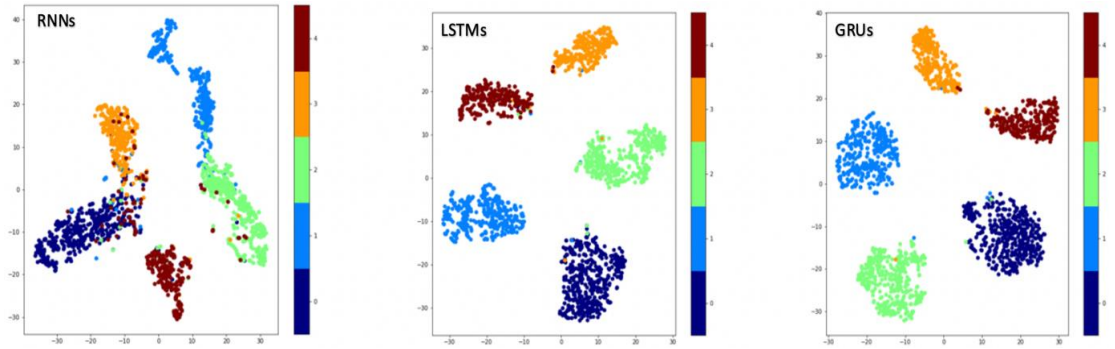

Figure 6.3 Visualization of Antisocial behaviour classes using t-SNE w.r.t GloVe embeddings (0- General/Non-ASB, 1- Failure to conform to lawful behaviour, 2- Irritability and aggressiveness, 3- Reckless disregard for safety, 4- Lack of Remorse)

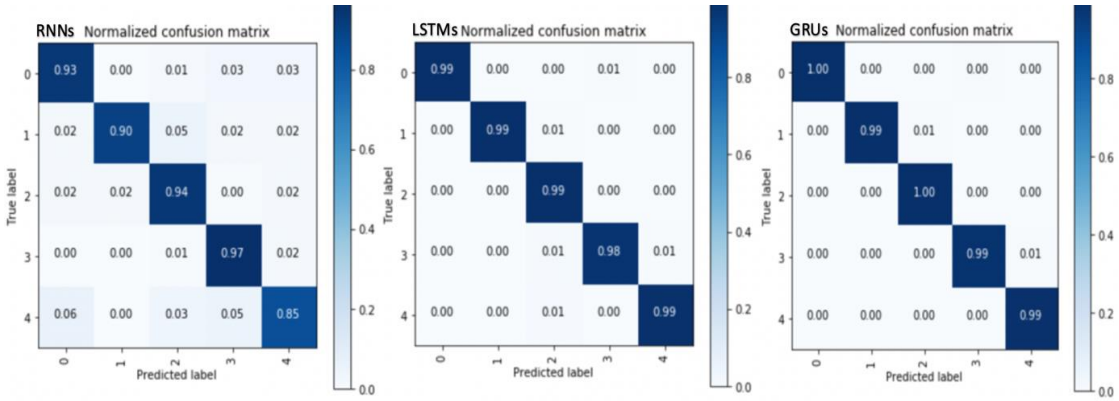

Figure 6.4 Confusion matrix. Deep Learning models w.r.t. GloVe embeddings (0- General/Non-ASB, 1- Failure to conform to lawful behaviour, 2- Irritability and aggressiveness, 3- Reckless disregard for safety, 4- Lack of Remorse)

To further understand the misclassifications by the experimental architectures (RNN, LSTM, and GRU) and to quantify the classification accuracy among classes, confusion matrices were generated and are presented in Figure 6.4. These confusion matrices provide finer-grained insight into the classification results. Since 3-fold cross-validation was used in this chapter, three sets of confusion matrices were generated for each architecture; each with its own

accuracy and misclassifications. To avoid any interpretation bias, the matrix with the highest accuracy and lowest misclassifications from the 3-fold were chosen for comparisons. RNN performance was inferior to the other architectures with the highest number of misclassifications. For Class 4 (*Lack of remorse*), only 85% of the tweets were correctly classified. Six percent were classified as non-antisocial, 5% as Class 3 (*Reckless disregard for safety*) and 3% as class 2 (*Irritability and aggressiveness*). Similarly, the algorithm did not perform well in classifying non-antisocial tweets, achieving an accuracy of 90%. Three percent were classified as Class 3 and another 3% as Class 4 tweets. It performed best with Class 3, with 97% accuracy.

In contrast, GRUs were able to classify all tweets with high accuracy. It achieved 100% accuracy with Class 0 and Class 2 tweets. Other three classes achieved 99% accuracy each. LSTMs performed better than RNNs with fewer misclassifications, nonetheless, they did not perform as well as GRUs. Most LSTM misclassifications were for Class 3 (*disregard for safety*). It can be inferred that Class 2 tweets were mostly classified correctly using all the algorithms and Class 4 experienced the most misclassifications, indicating that aggression can be depicted relatively correctly compared to lack of remorse.

6.4.6 Error Analyses

The section investigates some of the inaccurate classification outputs; the analysis of which can be leveraged to reannotate some of the tweets and retrain

algorithms for further accuracy improvement. The analysis was conducted to understand the source of misclassifications. The examples were generated using a Glove-GRU combination that produced the highest results among the feature-model combinations. Table 6.5 shows some of the misclassified tweets along with their actual and predicted label and Table 6.6 shows the example of correctly classified tweets.

Table 6.5 Misclassification example

ID	Post	Actual Label	Predicted Label	Probability
P_1	*next person that points at my shoes and says what are thoose im responding with go to hell ya*	*2*	*1*	*0.57160133*
P_2	*hey anytime fitness stick it up your ass you bunch of crooks this is been cancelled on paper by telephone you re damn right and cancelled by my credit card had to get a new credit card because of you crooks people stay away from anytime fitness scam*	*2*	*3*	*0.45120385*
P_3	*we don t need him to go to the cemetery to prove he s a patriot barack obama did every year and he was anything but the patriots put that in your pipe and smoke it or stick it up your ass*	*2*	*3*	*0.5395448*
P_4	*what would you do if i jumped from the building cry a bit and then forget*	*3*	*4*	*0.60184914*
P_5	*so what if i am speeding on a slipery road on a rainy day its is all fun and part of fun*	*3*	*2*	*0.68005776*
P_6	*i do not like wearing protective gears when i am working in the fields i am not worried about my life if will die i will just die no questions*	*3*	*4*	*0.5530549*
P_7	*i hope you die slow from the puncture of the blade mais um choker da ogoticario*	*4*	*2*	*0.64129925*
P_8	*i love pissing people off with political correctness lmao if you re going to be an arrogant bigot i m going to throw it in your face*	*4*	*2*	*0.5357206*
P_9	*i broke her tooth and threw her on the floor not ashamed at all she deserves it*	*4*	*3*	*0.5527889*
P_{10}	*2018 10 21 contentment breeds in our disintegration like bitter pills digested by the sick i wish you luck and hope you ve found your medicine*	*0*	*2*	*0.67318283*

From the output data analysis, it can be discerned that almost all tweets that were classified correctly were classified with a high probability index. However, all the misclassified tweets had a lower probability index. The algorithm had to pick the highest index among the set of low indexes pointing towards other labels. These tweets that classified with lower probability can be considered as borderline tweets. There must be some words, phrases or semantic characteristics in these tweets that pointed towards other labels, leading the model misclassification.

Post1's actual label is Class (*Irritability and aggressiveness*). The post does show a degree of irritability on part of the writer, however, was classified as Class 1 (*Failure to conform to social norms*). This might be due to the use of the word 'hell,' which is not desirable for use in this context. This really is a challenging classification task even by human standards. Similarly, P_4 can be considered another difficult post to classify due to its context. The post's actual label is Class 3 (*Reckless disregard for safety*), however, it classified wrongly as Class 4 (*Lack of remorse*). The post clearly displays disregard for safety when the writer suggests jumping from the building, however, at the same time it exhibits a subtle suggestion that the person to whom the post is referring to, may not show any remorse. There are several more examples presented in Table 6.5 and almost all of these are challenging to classify due to their confusing structure and word selection. These, along with other misclassified posts, can be used to retrain the model as a part of active learning.

Table 6.6 Correctly classified tweet examples

ID	Post	Actual Label	Predicted Label	Probability
P_1	*legally fuck the law laws are merely words on paper backed by the promise of violence if they are not obeyed words on paper have no authority over any of us just like imaginary lines have no authority*	*1*	*1*	*0.9983961*
P_2	*this is where the police and prisons need to point out who these ppl are. an eye for an eye should be the approach and these men shud be beaten day after day after day filthy animals shud not be protected by law*	*1*	*1*	*0.9969779*
P_3	*you bastard who the hell do you think you are you are supporting a national sovereignty breach and advocating open borders get the hell out of this country*	*2*	*2*	*0.98594415*
P_4	*we stand with saudi arabia my ass go be the president of a different country and suckle the teets of authoritarians you fucking asshole i personally like my journalists alive and in one piece*	*2*	*2*	*0.9549856*
P_5	*i took a selfie from my window i had one leg in the air and one on the balcony yes it was dangerous but you know what it was worth it now i have the best selfie and i am still alive bingo*	*3*	*3*	*0.9978516*
P_6	*once on read light i can speed like anything to get ahead of anyone i just like to be number one on the road and do not care how much faster i have to drive*	*3*	*3*	*0.99510854*
P_7	*i hope you die drowning in medical debt then this joke will be hillarious*	*4*	*4*	*0.99659914*
P_8	*you haven t known me long enough to know i love pissing people off she crosses her arms*	*4*	*4*	*0.95959044*
P_9	*happy birthday to young rebel star prabhas wish you luck to strike gold with sahoo fans are showering prabhas with well wishes a day before his birthday*	*0*	*0*	*0.99972147*
P_{10}	*a funny story while canvassing in md i was wearing my beto shirt lady opens door i say my speil she says i would never vote for beto cruz he s a fraud and i don t wish you luck a ya andb um md*	*0*	*0*	*0.91326326*

In Table 6.6, the second part of the classification output is presented. These are the tweets that were correctly classified despite containing confusing context and content. The high probability index implies a high degree of confidence in classifying. For instance, P_2 was correctly classified as Class 1 (*Failure to conform to social norms concerning lawful behaviour*), even though the post has some indication of irritability and aggressiveness (Class 2). Similarly, P_7 was correctly

classified as Class 4 (*Lack of remorse*) even though the post contains the word 'drowning', leaning more towards Class 3 (*Reckless disregard for safety*).

Word embeddings output for some of the most commonly occurring and decisive terms in the dataset was generated and are presented in Table 6.7. In terms of 'regret', mostly related words (e.g. 'confess', 'admit') and synonyms (e.g. 'remorse') were captured in the vector space, highlighting their interdependence. The word 'broke 'exhibits a more general meaning and, in contrast, attracted a wider range of words implying different meanings and contexts. Similar to the word 'regret' the word 'bastard' attracted related and synonyms taboo words. A few more examples are presented, and these afford the opportunity to have a deeper dive into the inner workings of the word embeddings.

Notwithstanding the syntactical linguistic variety that is inherited by any social media platform, the deep learning model utilizing word embeddings, exhibits its ability to distinguish relationships between concepts that are imperative to any NLP task. The wrongly classified tweets and their corresponding prediction probabilities are returned, enabling the identification of source classification confusion, leading to the potential refinement of the classes. Moreover, the analysis of the embeddings facilitates the opportunity for continuous performance improvement and active learning.

Table 6.7 Word embeddings

ASB Related Words	**Learnt by GloVe Embeddings**
Regret	*regret, remorse, shame, sadness, apologies, admit, doubt, pity, mistake, afraid, ashamed, sorrow, feelings, confess, worry, guilt, disaapointed, embarrassment, forgive, knowing.*
Suffering	*suffer, illness, misery, pain, endured, endure, sickness, overcome, struggle, agony, painful, grief, dying, pains, fear, illnesses*
Broke	*came, ended, fell, broken, finally, pulled, knocked, went, turned, kicked, blew, got, took, had, stoped dropped, threw, ran, pushed*
Painful	*pain, pains, discomfort, uncomforatble, painfully, painfull, frustating, ache, agony, embarrassing, suffering, endure, difficult, ordeal, suffer ,awful, horrible, terrible, stressful*
Scared	*afraid, terrified, worried, pissed, scary, shocked, scare, confused, angry, embarrassed, hell, crying, mad, hurt, scares, worry, anxious, tired, heck, crazy*
Kill	*killing, kils, killed, destroy, dead, enemy, attack, fight, hell, poison, killer, dying, steal, deadly, evil, him, death, hurt, revenge, let, murder, attacked*
Hate	*hating, hates, stupid, hated, despise, don't, crap, blame, hell, loathe, shit, afraid, hatred, idiots, complain, damn, shame, bother, dumb, cuz, bad*
Fight	*fight, fighting, battles, fought, battle, fought, battle, fighter, boxing, battling, combat, bout, opponent, punches, against, beating, defend, martial, defeat, revenge*
Bastard	*fucker, idiot, scumbag, motherfucker, moron, faggot, coward, bitch, asshole, dumbas, fuckin, shit, wanter, dumb, hell, stupid, arse, douchebad, buger, shithead*

The model demonstrates its robustness in gauging the subtle clues within the dataset, even with a relatively small training dataset (approx. 5500). The absence of substantial research work in behaviour studies on social platforms makes it a valued starting point in detecting and eliminating online ASB. The need for manual feature engineering efforts is eliminated due to the use of a deep learning architecture that facilitates a systematic and automated approach. Considering the extremely noisy characteristics of data collected from social media platforms, the ability to generate text representation that is impervious to irrelevant factors and highly precise and relevant to the crucial aspect of online ASB, is indispensable. The traditional machine learning approach of feature

extraction, also known as shallow processing, is limited in nature and works only with surface-level features and well-structured documents, however, it can some time prove inadequate for handling challenging user-generated data. Thus, more sophisticated techniques are imperative if subtle and often latent aspects are needed for accurate class assignment.

6.5 Summary of Findings

Online ASB is a public health threat and a social problem. It is a prevailing pattern of disregard for, and violation of, the rights of others. The exhibition of ASB might be an entertaining act for a perpetrator, nonetheless, it can lead a victim into anxiety, depression, low self-esteem, self-confinement, or suicidal ideation. Twitter and other online platforms can sometimes become incubators for such behaviour leading to numerous societal problems. Given the large amount of unstructured social media data, a scalable and robust automatic tweet classification technique is imperative for the efficient management of online content, and for the timely intervention by the platforms to prevent dire situations.

This chapter proposes an approach for multi-class identification of ASB from social media posts using the state-of-the-art deep learning models. Its main contributions are: 1) Benchmark medium-scale ASB dataset with multi-class annotation, 2) Development of a deep learning classification model performing successfully compared to different architectures, 3) Performance validation of

the deep learning model against traditional machine learning baselines, 4) Error analyses using visually enhanced graphical interpretation of similarities among the different classes of ASB to intercept the sources of misclassifications and 5) Knowledge discovery related to ASB tweets by leveraging descriptive analyses. An extensive set of experiments were conducted implementing different feature-model combinations of deep learning architectures, with the results presented in Table 6.4. Overall, the deep learning models with GloVe embedding achieved higher scores in all evaluation metrics compared to the same models using Word2Vec embeddings, and also to the traditional machine learning algorithms (except for RNNs). The highest accuracy was achieved by GRUs using GloVe embeddings with a batch size of 32 and Nadam optimizer. Along with the empirical validation of the superiority of the proposed deep learning model over the traditional machine learning algorithms, a realistic solution for the real-world problem of ASB has been presented. As a part of the study, a benchmark ASB corpus was created by manually annotating the tweets under the supervision of domain experts. Such corpora can reduce the time and cost needed to prepare ASB datasets for future studies. The important of this capability was emphasized in [4].

To better understand the inner working of the classification process and to analyze errors, the dimensionality reduction set of scatter plots were created to demonstrate both performance and misclassification in 2D space. To further quantify the classification scores of each class, confusion matrices were

generated to complement the analyses and pinpoint the main root causes of misclassifications. The matrices also provided decision support for choosing the optimal model for specific ASB category detection. For example, GRUs performed better when detecting Class 3 and Class 0, and LSTM and GRU performance was comparable in detecting the remaining three classes.

Despite the results achieved, the findings presented in this chapter should be considered in light of some limitations. The size of the benchmark dataset is moderate in nature (approx. 5500 tweets). This is due to the laborious process of manual annotation. Nevertheless, the tweet distribution among the five categories was quite similar, and the size of the corpus proved adequate for training and testing. Furthermore, the word embeddings technique inherently expands the feature vectors, essentially leveraging even a medium-size dataset. The advantages of data collection from other platforms, such as Facebook and Reddit, were also recognized. Analyses of other data sources could further enrich research in this area (class composition across different platforms). Moreover, the new categories of ASB can be identified by continued monitoring of social media discourse. Regardless of the limitations mentioned above, an approach towards proactively detecting and mitigating the detrimental impacts of ASB on the mental and physical health of victims, using cutting technology, has been proposed.

CHAPTER 7

CONCLUSION AND FUTURE WORK

The current chapter concludes the thesis by summarizing the main contributions of the research performed. It discusses some of the limitations of the study and highlights future directions that the study has laid the groundwork for. The chapter also briefly discusses the overall impact of the study within the computational social science research area.

7.1 Summary and Contributions

Social media is a relatively a new phenomenon that has emerged recently and has gained popularity and status as a 'place to be' for individuals of all age groups. It is a source of abundance for knowledge and entertainment. Individuals visit social media platforms for amusement and to socialize, and businesses use them to reach out to customers. No longer does a customer have to wait to hear back from a service representative of an organization. He/she can send a message directly to the management or even the CEO via one of the social media platforms. It works well the other way around as well; the CEO can talk directly to customers for feedback instead of relying on internal channels.

Social media has created new types of jobs and has enabled many to perform parts, if not all, of their jobs online. It is available everywhere as long as one has access to an Internet connection, which is quite ubiquitous; enabling individuals

to carry on a variety of tasks related to their daily lives. Therefore, social media has become an irreplaceable and valuable source of data and knowledge related to numerous pressing issues for human society. The continuous stream of user-generated data and its often involuntarily nature, along with the number of unrestricted avenues to collect it, has paved the way for the increasing popularity of the social media-based research to extract actionable knowledge. Through the active participation of users and its worldwide reach, social media platforms have demonstrated infinite potential for all types of behaviour studies and research.

Even with all their bells and whistles, online platforms bring their own costs. In the name of free speech, they have become breeding grounds for undesirable behaviours. Online antisocial behaviour is one such issue that has been the main focus of this thesis. It prevents fair participation on social media platforms and puts vulnerable groups of people at risk. Unacceptable behaviour in the form of threats, disregard for safety, lack of remorse, and failure to conform to lawful behaviour is afflicting online communities. This behaviour needs to be detected and eliminated, making online platforms safer places for all. Thus, the main objectives of this research study are as follows:

- *To develop an approach to extract and examine distinctive human behaviour patterns and personality traits of users from their use of different social media platforms*

- *To introduce a framework that can identify antisocial behaviour from online discourses using natural language processing techniques, traditional machine learning, and deep learning*
- *To propose a technique for automatic extraction and classification of different types of antisocial behaviours with high accuracy and to discover new knowledge related to such behaviours that can be utilized to combat its prevalence online.*

To achieve these objectives successfully, the research has proposed three benchmark datasets from two different online social media platforms: Twitter and the Swarm app. These were then annotated in consultation with, and under the supervision of, a domain expert who works in behaviour studies. Apart from high velocity and high volume, the informal, noisy, non-technical and descriptive characteristics of user-generated data creates further complications in extracting meaningful knowledge. The traditional machine learning algorithms delivered a sub-par performance compared to their deep learning counterparts, due to online data's feature sparsity and often non-semantic representation. For example, the terms mental abuse, mental assault, and behaviour violence, would all be treated as different features, even though they all share the similar meaning in the context of antisocial behaviour. As a result of this, the classification performance is negatively impacted. Traditional textual features are not typically generalizable, nor do they scale well given the large variety and complexity of expressions on the social media platforms. To effectively understand and handle the details of user-generated content on online platforms, an alternative approach is needed. One that is capable of

capturing precise semantic and syntactic relationships between words and sentences. State-of-the-art deep learning methods with word embeddings have, therefore, been utilized to address the shortcomings of traditional machine learning methods and related feature extraction techniques. To the best of our knowledge, this research work is the first attempt of empirical validation of deep learning methods on the real-world scenario of extracting online ASB and related information in the computational behaviour science domain. Furthermore, other feature extraction techniques that are commonly used in natural language processing and text mining tasks, such as bag-of-words (BoW) and psycholinguistic (LIWC), were also experimented with, to evaluate performance, gauge effectiveness and for performance comparison.

The first research step entailed establishing whether the behaviour and personality traits of an individual or group can be captured from their online activities (Chapter 4). One of the hypotheses for this research assumes that this is true, and was later conclusively supported by the findings presented in Chapter 4. First, a benchmark dataset from the Swarm app and the Twitter platform was created. The data collected from these platforms was cleaned, processed, and manually annotated to transform them into a gold standard. Tools and programming paradigms such as MeaningCloud, LIWC, and SPSS were implemented for various in-depth analyses and experimentation. Geo and spatial temporal analyses were conducted to study how people moved around in a city and at different times of the day, and on different days of the week. Text

mining in the form of language identification, topic extraction and text clustering were conducted to find out about the topics people were interested in at a particular point in time and their reaction/s to them. Sentiment analysis and psychometric analysis afforded the opportunity to understand users' feelings, thoughts, motivations and personalities. Descriptive statistical analyses complemented further behaviour analyses. The results and findings not only validated a lot of common knowledge and understandings related to human behaviour in behaviour science literature, but also contributed new discoveries. The findings are a testament to the reliance and validity of the proposed approach and framework offered by this thesis. Finding certain types of online behaviour traits also paved the way for experimentation with other crucial behaviour traits.

The value and the benefits of social media have been emphasized throughout this thesis, along with the unacceptable and antisocial behaviour that it brings. Thus, the need for such behaviour to be successfully captured, at scale, and classified was established. The need to utilize advance computational techniques in the form of machine learning and deep learning were also recognized.

Considering the problem statement and the associated limitations, the deep learning-based method for detecting online antisocial behaviour has been proposed in Chapter 5. First and foremost, a benchmark data set with labels ASB

and non-ASB was constructed. The problem of identifying online antisocial behaviour was treated as a binary text classification problem in which a particular post is categorized based on its textual content. Word embeddings and advance features were then extracted from the unstructured dataset to capture word to word dependencies. Apart from measuring the degree of similarity between the words, word embedding also accurately captured the vital association, relationship, and links between disparate concepts within the text. The pre-trained word embeddings, Word2Vec, was implemented for this task. Cutting-edge deep learning algorithms (GRU, LSTM, RNN, CNN) were subsequently experimented with, and the accuracy along with other evaluation metrics was compared. The efficiencies of the traditional machine learning algorithms (Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes) were also explored to provide a baseline for the deep learning techniques. For traditional machine learning algorithms, word frequency and TF-IDF features were applied. The empirical evidence demonstrated that the deep learning architectures outperformed the traditional machine approaches. Almost all deep learning architecture outperformed their traditional machine learning counterparts with CNN outperforming all. With regards to knowledge discovery, some of the most distinctive features between antisocial and non-antisocial categories were investigated for further awareness of the prevalence of such behaviours. Semantic coherence, in the form of word clouds, correlation between the frequently occurring words in both categories (ASB and non-ASB), and the heat map of word relationships were presented.

The proposed approach is critical in assisting online platforms, health care providers, policy makers and government organizations and other support groups combat online ASB and prevent its proliferation.

The aforementioned binary classification task was subsequently extended to a multi-class classification problem for further extracting fine-grained information related to distinctive forms of online ASB (Chapter 6). The automated behaviour categorization allows the online platform to swiftly address the issue of high-velocity and large volume data that is the core of any social media platform. The four most prevalent forms of online ASB were identified to assist online platforms to evaluate the nature and seriousness of the distinctive types of ASB. Certain type of ASBs, which can cause dire consequences, may need to be dealt with as an utmost priority compared to the other types. There are more than four types of ASBs, however, for this research, the most problematic were chosen. These four forms of ASB are failure to conform to social norms and disrespect for the law, aggressiveness and irritability, reckless disregard for the safety of others and self, and a lack of remorse after having mistreated and hurt someone. Analogous to the binary classification, an appropriate benchmark dataset (gold standard) was created. The cutting-edge deep learning architectures were experimented with and their relative performances were compared. The two most popular and widely used word embeddings namely, Word2Vec and GloVe, were implemented and their performances were evaluated on the benchmark dataset. Furthermore, a

comprehensive set of experiments covering different feature-algorithm combinations were carried out on the benchmark dataset to determine the most optimal feature-model combination. The empirical evaluation demonstrated that the state-of-the-art deep learning architectures combined with word embeddings yielded a higher performance (in all the commonly used evaluation metrics namely accuracy, precision, F1 and recall) compared to traditional machine learning architectures. The GRU+GloVe embedding combination achieved the maximum accuracy (approximately 99%). With the exception of Support Vector Machine, almost all the traditional machine learning algorithms, that were experimented with, yielded a sub-par performance compared to their deep learning counterparts. These traditional machine learning algorithms were also applied using the same word embeddings, namely GloVe and Word2Vec, empirically validating the superior performance of the state-of-the-art deep learning algorithms on the multi-class benchmark data set. Decision tree's performance was inferior to all the architectures tested in this research.

Bearing in mind the impact of hyperparameters on the performance of the deep learning algorithms, different combinations of optimizers, activation functions, batch sizes and the number of recurrent units were experimented with. The following combination resulted in the optimal result: Activation function-Relu, Optimizer-Nadam, Number of recurrent units-128, and Batch size-32. Visually enhanced interpretations of the classification process, in the form of scatter plots and confusion matrices, were generated for further analyses. Error analyses afforded the opportunity to reannotate some of the misclassified posts for

further performance enhancement. Fine-grained descriptive statistical analyses were performed on the multi-label annotated dataset for comparative analysis of distinctive forms of ASB classes, and to facilitate knowledge discovery from these four selected forms. All four types of posts varied in the number of words written, number of stop words used, most commonly occurring words, and the relationships between these words. It was interesting to find that people use short posts/sentences when exhibiting aggression and irritability and relatively lengthy posts when exhibiting lack of remorse.

7.2 Study Limitations

Comprehensive analysis and experiments related to various behaviour traits were conducted, and related contributions to the existing knowledge were made in this research. However, the study has some limitations, and these are listed below.

Size of datasets: The gold standard benchmark datasets that were constructed as a part of this research were considered moderate in size. This is due to the labour-intensive nature of manual annotation of tweets and posts. Collecting online data and manually labelling it was the only available choice due the lack of existing pre-defined lexicons and a publicly available benchmark dataset with fine grained annotations. Furthermore, to establish inter-rater reliability of the datasets, a critical step imperative to the high standard of annotation, added to the amount of time required to construct high quality datasets. A large number

of posts were initially collected however, many of these were discarded as they did not fit the criteria for this study. Since manual annotation requires a lot of resources, it was hard to get datasets any bigger than that used in the study. Automatic annotation, with some reliability index built in it, could have assisted in constructing a bigger dataset. Most of the data annotation in this area is taking place manually, limiting the ability to construct larger training and testing datasets. Nevertheless, the word embeddings and the deep learning techniques proved robust and successful in capturing the subtle and latent details which are often crucial for any text classification problems from medium-sized datasets.

Source of dataset: Data collected from social media platforms is considered as short text/message, often containing emojis, slang, spelling mistakes, and some characters from other human languages. For the analysis and experimentation of this study, the datasets were pre-processed only to remove duplicate posts and some emojis. The spelling, grammar and slang were not altered. There is a general consensus in the research community relating to the veracity of social media data, and hence the conclusion derived from it. Different platforms support different types of posts which may vary in content, length and accompanying media. The data for this research was mainly collected from Twitter and the Swarm app. If used with Twitter data, data from other platforms could cement the findings, further validate the deep learning classification performance, and certify the generalization of the framework. Nonetheless, this

is the first work in the domain of detecting online ASB, and the study lays the groundwork for future work in the area.

Manual annotation: Manual annotation of the posts was conducted under the supervision of a domain specialist. It is a cumbersome and subjective task. Even though DSM-5 [1] guidelines were used to identify ASB from online posts, some of the annotations might have been influenced by the rater's subjectivity. To combat this, inter-rater reliability was established using kappa coefficient. A lot of posts that did not meet the criteria were discarded. Despite the significant efforts to conduct precise annotation under the supervision of a domain expert, some discrepancies may have remained in the data set. Nonetheless, risk such as this forms an inherent part of the majority of machine learning training methods. Such minor annotation errors can impact negatively on the performance of a classifier even when all steps are taken, and every effort is put in to construct a gold standard dataset. The Kappa statistic did bring in a sufficient degree of agreement among annotators and hence minimized any risk of mis-annotation.

Technique novelty: The state-of-the-art machine learning and deep learning model were experimented with to detect and classify social traits and ASB from online corpora. One of the objectives of the research study was to empirically validate the superior classification performance of deep learning methods over the traditional machine learning algorithms, working with a real-world case

study. The deep learning algorithm's accuracy was improved by feature engineering using world embeddings, and fine tuning hyperparameter settings. Therefore, the novelty of the research work is considered as a practical and successful application of deep learning technology to a real-world scenario of detecting and classifying online ASB, for which a research gap was identified. A major part of the thesis therefore, focuses on the extraction of knowledge and actionable insights that online platforms, government organizations, healthcare providers, policy makers and support groups can utilize to combat online ASB, which is a major societal problem.

7.3 Future Research Directions

The research in this project can be extended in several directions, some of which are discussed below.

New theories: Social scientists and psychologists have studied ASB for a long time. Most of the existing theories related to this behaviour are based on real-world interactions, however online ASB is relatively a new phenomenon. Scientists have recently discovered this phenomenon with the surge in Internet usage in general, and social media in particular which took off in 2007. Some of the aetiologies and characteristics of online ASB may vary from the real-world exhibition of ASB. This research is one of the first to delve deeply into online ASB and its various manifestations. The exhibition of online ASB may lead to the formation of new ASB theories only relevant to the online version. One

example of this could be that in the real world, a perpetrator usually has power over a victim in the form of a large and strong physical appearance, money, status, position in society, job status, etc. which may not matter in an online scenario where a perpetrator and his/her aforementioned attributes are not necessarily visible to the victim. What that means is that a young boy with a small body structure, no money and status can bully a large-bodied individual with a lot of money and good status in society. The young boy's exhibition of ASB may not be influenced by the big guy's position in society, which is contrary to many ASB theories in the real world. This research project lays the ground work for many similar theories related to online ASB.

Other types of online antisocial behaviour: In this study, the focus was laid on the four predominantly occurring online ASBs. The study can be extended to the other categories such as deceitfulness, which is indicated by lying, repeated conning of others for pleasure and personal profit, and the use of aliases. Impulsivity, which can be considered as failure to plan ahead, is another ASB that demands attention. Furthermore, some other types of ASB that are worthy of consideration are consistent irresponsibility, failure to honour financial obligations, and unreliability. The types of antisocial behaviour that are not covered in this research cover all the remaining aspects of online ASB.

Other types of personality and behaviour traits*:* According to the DSM-5, which is a reference guide used by social scientists and psychologists to study

behaviour and personality traits/disorders, antisocial behaviour is one of the 10 personality disorders/traits. There are nine other traits that prevail online and can be studied in future work. These are:

- ***Paranoid personality*** which is a pattern of suspiciousness and distrust.
- ***Narcissistic personality*** which relates to a lack of empathy, need for admiration and patterns of grandiosity
- ***Schizoid personality*** which relates to restricted emotional expression and detachment from social relationships
- ***Avoidant personality*** which relates to hypersensitivity to negation, feelings of social inhibition and inadequacy
- ***Schizotypal personality*** which relates to perceptual and cognitive distortions, acute discomfort in relationships, and eccentric behaviour
- ***Dependent personality*** relates to excessive need to be looked after, and a pattern of clinging and submissive behaviour
- ***Borderline personality*** relates to impulsivity and is a pattern of instability in relationships
- ***Obsessive-compulsive personality*** relates to a pattern of perfectionism, control and orderliness
- ***Histrionic personality*** relates to attention seeking and excessive emotionality.

Apart from the above-mentioned behaviours, suicidal ideation, terror intentions, and the likelihood of causing community harm, are some of the other areas that can be explored with a methodology similar to the one used in this research project.

Demographic investigation: Posts that were collected as a part of this research project have associated metadata. This metadata has important information related to a user's age, gender, geographical location, interests, etc., which can be utilized to further investigate the victims and preparators of ASB. The information obtained can be utilized for profiling potential victims and perpetrators. These profiles can then be monitored on regular bases for a proactive approach to tackling online ASB. The approach can also be used for other personality and behaviour traits.

Dataset size: Collecting posts and manually annotating them is an intensive task that requires much time and resources. The datasets used in this study were moderate in size, bearing in mind the time and resources required to construct one. For further studies in the area, the size of datasets can be expanded by collecting more posts. The dataset can also be enhanced by collecting data from other platforms such as Reddit and Instagram. This can assist in finding validations and enabling fine-grained analyses. Having access to data from different platforms can also enable researchers to use data from one platform to train a model, and then test this model using data from a different platform. This

validation process can make sure that the framework and the model will work regardless of the platform on which they are implemented.

Geo-location correlation: Metadata in the form of geo-attributes that are associated with every post can afford an opportunity to do analyses based on the geographical locations of both the victims and perpetrators of online ASB. The available spatial information can allow for studies on correlation between areas with higher numbers of perpetrators and factors such as poverty, education and population density of that area. Correlation between political variants, weather, culture and religion with ASB can also form the basis of future studies.

Subcategories of victims: The proposed approach has studied all types of online ASB victims regardless of their age, gender, geographic location, sexual orientation, religion and political orientation. The similar is true for perpetrators. For future work, the study can be expanded to analyze victims and perpetrators based on all the aforementioned factors separately. For instance, it will be quite useful to know whether LGBT communities are targeted more with online ASB than the other groups of our society. Similarly, it will be useful to know whether males, and specifically Caucasian males, show more online ASB towards females in general, and females of colour in particular. Most of the details required for such studies can be obtained from the metadata of posts and from usernames. From the prevention point-of-view, studies based on these

factors can assist in the further development of profiles of potential victims and perpetrators, based on the newly formed theories. Therefore, appropriate proactive and constant monitoring can be used to avoid dire situations from occurring online.

Technique novelty: Further improvements to the proposed framework and methodology can be realized regardless of the high accuracies achieved using the state-of-the-art deep learning models. Deep learning approaches rely heavily on hyperparameter settings and these can be further optimized to achieve improved results. There is also a scope to establish automated post annotation techniques with high accuracy and reliability.

7.4 Final Remarks

Text mining reveals crucial topics, concepts and other meaningful information from large collections of social media data and other natural language resources. The technique enables algorithms and machine learning architectures in identifying latent facts and relationships between words, sentences, paragraphs and documents, utilizing numerous available feature extraction approaches. The research reported in thesis investigated various challenges posed by the upsurge in the use of complicated online social media networks and the related undesirable and antisocial personality and behaviour traits and patterns exhibited by its millions of users. The thesis presented a novel approach to study online behaviours and to combat those that are undesirable.

The research outcomes of this thesis are instrumental and far-reaching in promoting innovation and research of extensive text mining, NLP, machine learning, and deep learning techniques in computational behaviour science research. Most of human day-to-day activities have moved online and this is where personalities and behaviours are exhibited. Online social media has a big role to play, and is responsible for most of the online activities. Along with all the value that these platforms bring, they have become breeding grounds for unacceptable social behaviour. These unacceptable or antisocial behaviours can be detected and eliminated with the help of natural language processing and deep learning techniques. The thesis has successfully implemented these techniques and has laid the groundwork for future work in this crucial area of computational behaviour science research using cutting-edge technologies.

BIBLIOGRAPHY

[1] A. P. Association, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[2] X. Chen *et al.*, "Mining patients' narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate," *Frontiers in pharmacology,* vol. 9, p. 541, 2018.

[3] A. Nikfarjam, A. Sarker, K. O'connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association,* vol. 22, no. 3, pp. 671-681, 2015.

[4] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *Journal of biomedical informatics,* vol. 53, pp. 196-207, 2015.

[5] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 115-124.

[6] M.-A. Abbasi, S.-K. Chai, H. Liu, and K. Sagoo, "Real-world behavior analysis through a social media lens," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 2012: Springer, pp. 18-26.

[7] R. M. Chang, R. J. Kauffman, and Y. Kwon, "Understanding the paradigm shift to computational social science in the presence of big data," *Decision Support Systems,* vol. 63, pp. 67-80, 2014.

[8] H. Yin, B. Cui, Z. Huang, W. Wang, X. Wu, and X. Zhou, "Joint modeling of users' interests and mobility patterns for point-of-interest recommendation," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015: ACM, pp. 819-822.

[9] M. Sklar, B. Shaw, and A. Hogue, "Recommending interesting events in real-time with foursquare check-ins," in *Proceedings of the sixth ACM conference on Recommender systems*, 2012: ACM, pp. 311-312.

[10] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha, and S. Yenduri, "Application of Deep Learning to Sentiment Analysis for recommender

system on cloud," in *Computer, Information and Telecommunication Systems (CITS), 2017 International Conference on*, 2017: IEEE, pp. 93-97.

[11] P. V. Krishna, S. Misra, D. Joshi, and M. S. Obaidat, "Learning automata based sentiment analysis for recommender system on cloud," in *Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on*, 2013: IEEE, pp. 1-5.

[12] M. Shepard and A. Greenfield, "Urban computing and its discontents," *New York: The Architectural League of New York, New York USA,* 2007.

[13] A. Crooks *et al.*, "Crowdsourcing urban form and function," *International Journal of Geographical Information Science,* vol. 29, no. 5, pp. 720-741, 2015.

[14] E. Currid and S. Williams, "The geography of buzz: art, culture and the social milieu in Los Angeles and New York," *Journal of Economic Geography,* vol. 10, no. 3, pp. 423-451, 2010.

[15] N. Hochman and L. Manovich, "Zooming into an Instagram City: Reading the local through social media," *First Monday,* vol. 18, no. 7, 2013.

[16] T. Shelton, A. Poorthuis, and M. Zook, "Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information," *Landscape and Urban Planning,* vol. 142, pp. 198-211, 2015.

[17] E. Spyrou and P. Mylonas, "An overview of Flickr challenges and research opportunities," in *Semantic and Social Media Adaptation and Personalization (SMAP), 2014 9th International Workshop on*, 2014: IEEE, pp. 88-93.

[18] P. Brindley, J. Goulding, and M. L. Wilson, "A data driven approach to mapping urban neighbourhoods," in *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2014: ACM, pp. 437-440.

[19] L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using Flickr tags to describe city cores," *Journal of Spatial Information Science,* vol. 2010, no. 1, pp. 21-48, 2010.

[20] L. Pei *et al.*, "Human behavior cognition using smartphone sensors," *Sensors,* vol. 13, no. 2, pp. 1402-1424, 2013.

[21] M. Han, J. H. Bang, C. Nugent, S. McClean, and S. Lee, "A lightweight hierarchical activity recognition framework using smartphone sensors," *Sensors,* vol. 14, no. 9, pp. 16181-16195, 2014.

[22] J. Zhao, T. Wang, X. Xu, and Y. Yang, "Personalized LBSN Recommendation System," in *Proceedings of the 2017 International Conference on Management Engineering, Software Engineering and Service Sciences*, 2017: ACM, pp. 119-123.

[23] J. Melià-Seguí, R. Zhang, E. Bart, B. Price, and O. Brdiczka, "Activity duration analysis for context-aware services using foursquare check-ins," in *Proceedings of the 2012 international workshop on Self-aware internet of things*, 2012: ACM, pp. 13-18.

[24] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, "A survey on sentiment analysis and opinion mining for social multimedia," *Multimedia Tools and Applications,* vol. 78, no. 6, pp. 6939-6967, 2019.

[25] W. Zhang, C. Yu, and W. Meng, "Opinion retrieval from blogs," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 831-840.

[26] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070,* 2002.

[27] U. R. Hodeghatta, "Sentiment analysis of Hollywood movies on Twitter," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 2013: IEEE, pp. 1401-1404.

[28] A. Amolik, N. Jivane, M. Bhandari, and M. Venkatesan, "Twitter sentiment analysis of movie reviews using machine learning techniques," *international Journal of Engineering and Technology,* vol. 7, no. 6, pp. 1-7, 2016.

[29] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *LREc*, 2010, vol. 10, no. 2010, pp. 1320-1326.

[30] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *2014 Seventh International Conference on Contemporary Computing (IC3)*, 2014: IEEE, pp. 437-442.

[31] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, "Sentiment analysis on social media," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012: IEEE, pp. 919-926.

[32] Y. Yuan and Y. Zhou, "Twitter sentiment analysis with recursive neural networks," *CS224D course projects,* 2015.

[33] Y. Ren, Y. Zhang, M. Zhang, and D. Ji, "Context-sensitive twitter sentiment classification using neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016, vol. 30, no. 1.

[34] N. Collier and S. Doan, "Syndromic classification of twitter messages," in *International Conference on Electronic Healthcare*, 2011: Springer, pp. 186-195.

[35] M. Paul and M. Dredze, "A Model for Mining Public Health Topics from Twitter (Technical Report)," *Johns Hopkins University,* 2011.

[36] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using Twitter," in *Proceedings of the 2011 Conference on empirical methods in natural language processing*, 2011, pp. 1568-1576.

[37] G. Ramya and P. B. Sivakumar, "Advocacy monitoring of women and children health through social data," *Indian Journal of Science and Technology,* vol. 9, no. 6, pp. 1-6, 2016.

[38] G. Gkotsis *et al.*, "Characterisation of mental health conditions in social media using Informed Deep Learning," *Scientific reports,* vol. 7, no. 1, pp. 1-11, 2017.

[39] A. Acar and Y. Muraki, "Twitter for crisis communication: lessons learned from Japan's tsunami disaster," *International journal of web based communities,* vol. 7, no. 3, pp. 392-402, 2011.

[40] C. Caragea, A. Silvescu, and A. H. Tapia, "Identifying informative messages in disaster events using convolutional neural networks," in *International conference on information systems for crisis response and management*, 2016, pp. 137-147.

[41] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the crowdsourcing power of social media for disaster relief," *IEEE Intelligent Systems,* vol. 26, no. 3, pp. 10-14, 2011.

[42] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response," in *Proceedings of the 23rd international conference on world wide web*, 2014, pp. 159-162.

[43] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting information nuggets from disaster-Related messages in social media," in *Iscram*, 2013.

[44] D. T. Nguyen, S. Joty, M. Imran, H. Sajjad, and P. Mitra, "Applications of online deep learning for crisis response using social media information," *arXiv preprint arXiv:1610.01030,* 2016.

[45] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1474-1477.

[46] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science,* vol. 2, no. 1, pp. 1-8, 2011.

[47] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, and T. Graepel, "Manifestations of user personality in website choice and behaviour on online social networks," *Machine learning,* vol. 95, no. 3, pp. 357-380, 2014.

[48] L. Li, A. Li, B. Hao, Z. Guan, and T. Zhu, "Predicting active users' personality based on micro-blogging behaviors," *PloS one,* vol. 9, no. 1, p. e84997, 2014.

[49] C. Stavros, M. D. Meng, K. Westberg, and F. Farrelly, "Understanding fan motivation for interacting on social media," *Sport Management Review,* vol. 17, no. 4, pp. 455-469, 2014.

[50] C. Dijkmans, P. Kerkhof, and C. J. Beukeboom, "A stage to engage: Social media use and corporate reputation," *Tourism Management,* vol. 47, pp. 58-67, 2015.

[51] L. Dessart, C. Veloutsou, and A. Morgan-Thomas, "Consumer engagement in online brand communities: a social media perspective," *Journal of Product & Brand Management,* vol. 24, no. 1, pp. 28-42, 2015.

[52] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences,* vol. 110, no. 15, pp. 5802-5805, 2013.

[53] W. Fan and M. D. Gordon, "The power of social media analytics," *Communications of the ACM,* vol. 57, no. 6, pp. 74-81, 2014.

[54] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management,* vol. 35, no. 2, pp. 137-144, 2015.

[55] E. Kalampokis, E. Tambouris, and K. Tarabanis, "Understanding the predictive power of social media," *Internet Research,* vol. 23, no. 5, pp. 544-559, 2013.

[56] S. Asur and B. A. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 2010, vol. 1: IEEE, pp. 492-499.

[57] CNBC. https://www.cnbc.com/2016/02/02/chipotle-reports-fourth-quarter-earnings.html (accessed 27th October, 2017).

[58] Business Insider. https://www.businessinsider.com.au/foursquare-data-predicted-chipotle-results-2016-4?r=US&IR=T (accessed 27th October 2017, 2017).

[59] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," *ICWSM,* vol. 2011, pp. 81-88, 2011.

[60] J. Cranshaw, R. Schwartz, J. I. Hong, and N. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city," 2012.

[61] P. Aliandu, "Sentiment Analysis to Determine Accommodation, Shopping and Culinary Location on Foursquare in Kupang City," *Procedia Computer Science,* vol. 72, pp. 300-305, 2015.

[62] S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, 2013: ACM, p. 6.

[63] X. Zhou and L. Zhang, "Crowdsourcing functions of the living city from Twitter and Foursquare data," *Cartography and Geographic Information Science,* vol. 43, no. 5, pp. 393-404, 2016.

[64] L. Böcker, M. Dijst, and J. Prillwitz, "Impact of everyday weather on individual daily travel behaviours in perspective: a literature review," *Transport reviews,* vol. 33, no. 1, pp. 71-91, 2013.

[65] S. S. Sundar, *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative, 2008.

[66] G. E. Gigerenzer, R. E. Hertwig, and T. E. Pachur, *Heuristics: The foundations of adaptive behavior*. Oxford University Press, 2011.

[67] K. M. Griffiths and H. Christensen, "Website quality indicators for consumers," *Journal of medical Internet research,* vol. 7, no. 5, p. e55, 2005.

[68] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter,* vol. 19, no. 1, pp. 22-36, 2017.

[69] P. Lorenz-Spreen, S. Lewandowsky, C. R. Sunstein, and R. Hertwig, "How behavioural sciences can promote truth, autonomy and democratic discourse online," *Nature human behaviour,* vol. 4, no. 11, pp. 1102-1109, 2020.

[70] Z. Hilvert-Bruce and J. T. Neill, "I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming," *Computers in Human Behavior,* vol. 102, pp. 303-311, 2020.

[71] A. Chalk, "Researchers find that female PC gamers outnumber males," *PC Gamer,* vol. 28, 2014.

[72] S. Chess and A. Shaw, "A conspiracy of fishes, or, how we learned to stop worrying about# GamerGate and embrace hegemonic masculinity," *Journal of Broadcasting & Electronic Media,* vol. 59, no. 1, pp. 208-220, 2015.

[73] M. Popovac and P. Fine, "An intervention using the Information-Motivation-Behavioral Skills Model: Tackling cyberaggression and cyberbullying in South African adolescents," in *Reducing Cyberbullying in Schools*: Elsevier, 2018, pp. 225-244.

[74] J. H. Kuznekoff and L. M. Rose, "Communication in multiplayer gaming: Examining player responses to gender cues," *New Media & Society,* vol. 15, no. 4, pp. 541-556, 2013.

[75] M. E. Ballard and K. M. Welch, "Virtual warfare: Cyberbullying and cyber-victimization in MMOG play," *Games and culture,* vol. 12, no. 5, pp. 466-491, 2017.

[76] K. L. Gray, "Deviant bodies, stigmatized identities, and racist acts: Examining the experiences of African-American gamers in Xbox Live," *New Review of Hypermedia and Multimedia,* vol. 18, no. 4, pp. 261-276, 2012.

[77] A. Brighi *et al.*, "Self-esteem and loneliness in relation to cyberbullying in three European countries," *Cyberbullying in the global playground: Research from international perspectives,* pp. 32-56, 2012.

[78] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of suicide research,* vol. 14, no. 3, pp. 206-221, 2010.

[79] A. M. Gard, H. L. Dotterer, and L. W. Hyde, "Genetic influences on antisocial behavior: recent advances and future directions," *Current opinion in psychology,* 2018.

[80] E. Flouri and S. Ioakeimidi, "Maternal depressive symptoms in childhood and risky behaviours in early adolescence," *European child & adolescent psychiatry,* vol. 27, no. 3, pp. 301-308, 2018.

[81] M. Woeckener *et al.*, "Parental rejection and antisocial behavior: the moderating role of testosterone," *Journal of Criminal Psychology,* 2018.

[82] W. M. McGuigan, J. A. Luchette, and R. Atterholt, "Physical neglect in childhood as a predictor of violent behavior in adolescent males," *Child abuse & neglect,* vol. 79, pp. 395-400, 2018.

[83] D. B. Jackson, "The link between poor quality nutrition and childhood antisocial behavior: A genetically informative analysis," *Journal of Criminal Justice,* vol. 44, pp. 13-20, 2016.

[84] A. R. Baskin-Sommers, "Dissecting antisocial behavior: The impact of neural, genetic, and environmental factors," *Clinical Psychological Science,* vol. 4, no. 3, pp. 500-510, 2016.

[85] S. B. Manuck and J. M. McCaffery, "Gene-environment interaction," *Annual review of psychology,* vol. 65, pp. 41-70, 2014.

[86] L. W. Hyde *et al.*, "Heritable and nonheritable pathways to early callous-unemotional behaviors," *American Journal of Psychiatry,* vol. 173, no. 9, pp. 903-910, 2016.

[87] J. M. Beyers, R. Loeber, P.-O. H. Wikström, and M. Stouthamer-Loeber, "What predicts adolescent violence in better-off neighborhoods?," *Journal of Abnormal Child Psychology,* vol. 29, no. 5, pp. 369-381, 2001.

[88] D. L. Haynie, E. Silver, and B. Teasdale, "Neighborhood characteristics, peer networks, and adolescent violence," *Journal of Quantitative Criminology,* vol. 22, no. 2, pp. 147-169, 2006.

[89] T. Braga, O. Cunha, and Â. Maia, "The enduring effect of maltreatment on antisocial behavior: A meta-analysis of longitudinal studies," *Aggression and violent behavior,* 2018.

[90] V. J. Bland and I. Lambie, "Does childhood neglect contribute to violent behavior in adulthood? A review of possible links," *Clinical psychology review,* 2018.

[91] E. Anderson, "The code of the streets," *Atlantic monthly,* vol. 273, no. 5, pp. 81-94, 1994.

[92] E. Aisenberg and T. Herrenkohl, "Community violence in context: Risk and resilience in children and families," *Journal of interpersonal violence,* vol. 23, no. 3, pp. 296-315, 2008.

[93] D. Baskin and I. Sommers, "Exposure to community violence and trajectories of violent offending," *Youth violence and juvenile justice,* vol. 12, no. 4, pp. 367-385, 2014.

[94] S. Javdani, J. Abdul-Adil, L. Suarez, S. R. Nichols, and A. D. Farmer, "Gender differences in the effects of community violence on mental health outcomes in a sample of low-income youth receiving psychiatric care," *American journal of community psychology,* vol. 53, no. 3-4, pp. 235-248, 2014.

[95] E. R. Kimonis, L. C. Centifanti, J. L. Allen, and P. J. Frick, "Reciprocal influences between negative life events and callous-unemotional traits," *Journal of abnormal child psychology,* vol. 42, no. 8, pp. 1287-1298, 2014.

[96] Z. Walsh *et al.*, "Socioeconomic-status and mental health in a personality disorder sample: The importance of neighborhood factors," *Journal of personality disorders,* vol. 27, no. 6, pp. 820-831, 2013.

[97] L. S. Wakschlag, K. E. Pickett, E. Cook Jr, N. L. Benowitz, and B. L. Leventhal, "Maternal smoking during pregnancy and severe antisocial behavior in offspring: a review," *American journal of public health,* vol. 92, no. 6, pp. 966-974, 2002.

[98] P. A. Brennan, E. R. Grekin, and S. A. Mednick, "Maternal smoking during pregnancy and adult male criminal outcomes," *Archives of general psychiatry,* vol. 56, no. 3, pp. 215-219, 1999.

[99] D. M. Fergusson, L. J. Woodward, and L. J. Horwood, "Maternal smoking during pregnancy and psychiatric adjustment in late adolescence," *Archives of general psychiatry,* vol. 55, no. 8, pp. 721-727, 1998.

[100] C. L. Gibson and S. G. Tibbetts, "Interaction between maternal cigarette smoking and Apgar scores in predicting offending behavior," *Psychological Reports,* vol. 83, no. 2, pp. 579-586, 1998.

[101] P. Räsänen, H. Hakko, M. Isohanni, S. Hodgins, M.-R. Järvelin, and J. Tiihonen, "Maternal smoking during pregnancy and risk of criminal behavior among adult male offspring in the Northern Finland 1966 Birth Cohort," *American Journal of Psychiatry,* vol. 156, no. 6, pp. 857-862, 1999.

[102] L. S. Wakschlag and S. L. Hans, "Maternal smoking during pregnancy and conduct problems in high-risk youth: a developmental framework," *Development and psychopathology,* vol. 14, no. 2, pp. 351-369, 2002.

[103] L. S. Wakschlag, B. B. Lahey, R. Loeber, S. M. Green, R. A. Gordon, and B. L. Leventhal, "Maternal smoking during pregnancy and the risk of conduct disorder in boys," *Archives of general psychiatry,* vol. 54, no. 7, pp. 670-676, 1997.

[104] M. M. Weissman, V. Warner, P. J. Wickramaratne, and D. B. Kandel, "Maternal smoking during pregnancy and psychopathology in offspring followed to adulthood," *Journal of the American Academy of Child & Adolescent Psychiatry,* vol. 38, no. 7, pp. 892-899, 1999.

[105] C. E. Herbison *et al.*, "Low intake of B-vitamins is associated with poor adolescent mental health and behaviour," *Preventive medicine,* vol. 55, no. 6, pp. 634-638, 2012.

[106] A. Binns, "DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities," *Journalism Practice,* vol. 6, no. 4, pp. 547-562, 2012. [Online]. Available: https://www.tandfonline.com/doi/pdf/10.1080/17512786.2011.648988?needAccess=true.

[107] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, "Trolls just want to have fun," *Personality and individual Differences,* vol. 67, pp. 97-102, 2014.

[108] S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab, "Searching for safety online: Managing" trolling" in a feminist forum," *The information society,* vol. 18, no. 5, pp. 371-384, 2002.

[109] P. Shachaf and N. Hara, "Beyond vandalism: Wikipedia trolls," *Journal of Information Science,* vol. 36, no. 3, pp. 357-370, 2010.

[110] E. E. Buckels, P. D. Trapnell, T. Andjelovic, and D. L. Paulhus, "Internet Trolling and Everyday Sadism: Parallel Effects on Pain Perception and Moral Judgment," *Journal of personality,* 2018.

[111] R. B. Cialdini and N. J. Goldstein, "Social influence: Compliance and conformity," *Annu. Rev. Psychol.,* vol. 55, pp. 591-621, 2004.

[112] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *CSCW: proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, 2017, vol. 2017: NIH Public Access, p. 1217.

[113] N. Sest and E. March, "Constructing the cyber-troll: Psychopathy, sadism, and empathy," *Personality and Individual Differences,* vol. 119, pp. 69-72, 2017.

[114] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial Behavior in Online Discussion Communities," in *Icwsm*, 2015, pp. 61-70.

[115] S. Park, E.-Y. Na, and E.-m. Kim, "The relationship between online activities, netiquette and cyberbullying," *Children and youth services review,* vol. 42, pp. 74-81, 2014.

[116] C. Hardaker, "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions," ed: Walter de Gruyter GmbH & Co. KG, 2010.

[117] C. Chelmis, D.-S. Zois, and M. Yao, "Mining patterns of cyberbullying on twitter," in *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, 2017: IEEE, pp. 126-133.

[118] M. C. McHugh, S. L. Saperstein, and R. S. Gold, "OMG U# Cyberbully! An exploration of public discourse about cyberbullying on twitter," *Health Education & Behavior,* p. 1090198118788610, 2018.

[119] P. Lee, "Expanding the Schoolhouse Gate: Public Schools (K-12) and the Regulation of Cyberbullying," *Utah L. Rev.,* p. 831, 2016.

[120] J. Guberman and L. Hemphill, "Challenges in modifying existing scales for detecting harassment in individual tweets," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[121] A. G. Shirbhate and S. N. Deshmukh, "Feature extraction for sentiment classification on twitter data," *International Journal of Science and Research (IJSR),* vol. 5, no. 2, pp. 2183-2189, 2016.

[122] P. Zhao, X. Li, and K. Wang, "Feature extraction from micro-blogs for comparison of products and services," in *International Conference on Web Information Systems Engineering*, 2013: Springer, pp. 82-91.

[123] A. Stavrianou, C. Brun, T. Silander, and C. Roux, "NLP-based feature extraction for automated tweet classification," *Interactions between Data Mining and Natural Language Processing,* vol. 145, 2014.

[124] L. Yuan, "Improvement for the automatic part-of-speech tagging based on hidden Markov model," in *2010 2nd International Conference on Signal Processing Systems*, 2010, vol. 1: IEEE, pp. V1-744-V1-747.

[125] G. Forman and E. Kirshenbaum, "Extremely fast text feature extraction for classification and indexing," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 1221-1230.

[126] H. Saif, M. Fernandez, Y. He, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," 2014.

[127] H. Jadhao, D. J. Aghav, and A. Vegiraju, "Semantic tool for analysing unstructured data," *International Journal of Scientific & Engineering Research,* vol. 3, no. 8, 2012.

[128] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *LREC*, 2006, vol. 6: Citeseer, pp. 417-422.

[129] C. Strapparava and A. Valitutti, "Wordnet affect: an affective extension of wordnet," in *Lrec*, 2004, vol. 4, no. 1083-1086: Citeseer, p. 40.

[130] E. Montañés, J. Fernández, I. Díaz, E. F. Combarro, and J. Ranilla, "Measures of rule quality for feature selection in text categorization," in *international Symposium on Intelligent data analysis*, 2003: Springer, pp. 589-598.

[131] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 27, no. 8, pp. 1226-1238, 2005.

[132] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine learning research,* vol. 5, no. 9, 2004.

[133] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *Aaai*, 2006, vol. 6, no. 2006, pp. 775-780.

[134] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003, vol. 242, no. 1: Citeseer, pp. 29-48.

[135] V. Mazzonello, S. Gaglio, A. Augello, and G. Pilato, "A study on classification methods applied to sentiment analysis," in *2013 IEEE Seventh International Conference on Semantic Computing*, 2013: IEEE, pp. 426-431.

[136] A. Ritter, S. Clark, and O. Etzioni, "Named entity recognition in tweets: an experimental study," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 1524-1534.

[137] F. C. T. Chua, W. W. Cohen, J. Betteridge, and E.-P. Lim, "Community-based classification of noun phrases in twitter," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 1702-1706.

[138] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data," in *Proceedings of the international conference recent advances in natural language processing ranlp 2013*, 2013, pp. 198-206.

[139] T. Shanmugapriya and P. Kiruthika, "Survey on web content, mining and its tools," *International Journal of Science, Engineering and Research (IJSER) Volume,* vol. 2, 2014.

[140] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM,* vol. 56, no. 4, pp. 82-89, 2013.

[141] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[142] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, 1998: Springer, pp. 137-142.

[143] G. Biau and E. Scornet, "A random forest guided tour," *Test,* vol. 25, no. 2, pp. 197-227, 2016.

[144] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[145] C. Apté, F. Damerau, and S. M. Weiss, "Automated learning of decision rules for text categorization," *ACM Transactions on Information Systems (TOIS),* vol. 12, no. 3, pp. 233-251, 1994.

[146] J. R. Quinlan, "Simplifying decision trees," *International journal of man-machine studies,* vol. 27, no. 3, pp. 221-234, 1987.

[147] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, no. 1, pp. 81-106, 1986.

[148] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41-46.

[149] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, 1998, vol. 752, no. 1: Citeseer, pp. 41-48.

[150] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[151] S. Dara and P. Tumma, "Feature extraction by using deep learning: A survey," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018: IEEE, pp. 1795-1801.

[152] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, 2017: Ieee, pp. 1-6.

[153] D. Mandic and J. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. Wiley, 2001.

[154] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks,* vol. 5, no. 2, pp. 157-166, 1994.

[155] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, 2017: IEEE, pp. 1597-1600.

[156] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

[157] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[158] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.

[159] B. T. van Zanten, D. B. Van Berkel, R. K. Meentemeyer, J. W. Smith, K. F. Tieskens, and P. H. Verburg, "Continental-scale quantification of landscape values using social media data," *Proceedings of the National Academy of Sciences,* p. 201614158, 2016.

[160] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *arXiv preprint arXiv:0806.1256,* 2008.

[161] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science,* vol. 327, no. 5968, pp. 1018-1021, 2010.

[162] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modeling the scaling properties of human mobility," *arXiv preprint arXiv:1010.0436,* 2010.

[163] T. H. Silva, P. O. V. de Melo, J. M. Almeida, M. Musolesi, and A. A. Loureiro, "You Are What You Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food and Drink Habits in Foursquare," in *ICWSM*, 2014.

[164] G. B. Colombo, M. J. Chorley, M. J. Williams, S. M. Allen, and R. M. Whitaker, "You are where you eat: Foursquare checkins as indicators of human mobility and behaviour," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, 2012: IEEE, pp. 217-222.

[165] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in *Proceedings of the 21st international conference on World Wide Web*, 2012: ACM, pp. 519-528.

[166] V. Mayer-Schönberger and K. Cukier, *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[167] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, "I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011: ACM, pp. 2409-2418.

[168] A. Whiting and D. Williams, "Why people use social media: a uses and gratifications approach," *Qualitative Market Research: An International Journal,* vol. 16, no. 4, pp. 362-369, 2013.

[169] B. L. Fossen and D. A. Schweidel, "Social TV: How Social Media Activity Interacts With TV Advertising," *GfK Marketing Intelligence Review,* vol. 9, no. 2, pp. 31-36, 2017.

[170] C. Oh and S. Yergeau, "Social capital, social media, and TV ratings," *International Journal of Business Information Systems,* vol. 24, no. 2, pp. 242-260, 2017.

[171] D. Ruths and J. Pfeffer, "Social media for large studies of behavior," *Science,* vol. 346, no. 6213, pp. 1063-1064, 2014.

[172] K. Weller and M. Strohmaier, "Social media in academia: How the social web is changing academic practice and becoming a new source for research data," *IT-Information Technology,* vol. 56, no. 5, pp. 203-206, 2014.

[173] M. A. Zook and M. Graham, "Mapping DigiPlace: geocoded Internet data and the representation of place," *Environment and Planning B: Planning and Design,* vol. 34, no. 3, pp. 466-482, 2007.

[174] C. C. Aggarwal, *Data mining : the textbook*. Cham : Springer, 2015., 2015.

[175] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[176] N. Japkowicz and J. Stefanowski, *Big Data Analysis: New Algorithms for a New Society*. Springer, 2016.

[177] R. Kitchin, "Big Data, new epistemologies and paradigm shifts," *Big Data & Society,* vol. 1, no. 1, p. 2053951714528481, 2014/07/10 2014, doi: 10.1177/2053951714528481.

[178] J. Y. Tsai, P. Kelley, P. Drielsma, L. F. Cranor, J. Hong, and N. Sadeh, "Who's viewed you?: the impact of feedback in a mobile location-sharing application," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009: ACM, pp. 2003-2012.

[179] Foursquare. https://foursquare.com/about (accessed 23rd October 2017, 2017).

[180] Foursquare. https://developer.foursquare.com/docs (accessed 23rd October 2017, 2017).

[181] J. Frith, "Communicating Through Location: The Understood Meaning of the Foursquare Check-In," *Journal of Computer-Mediated Communication,* vol. 19, no. 4, pp. 890-905, 2014.

[182] Y. Sun, "Investigating "locality" of intra-urban spatial interactions in New York city using foursquare data," *ISPRS International Journal of Geo-Information,* vol. 5, no. 4, p. 43, 2016.

[183] Y. Zhang and M. Pennacchiotti, "Predicting purchase behaviors from social media," in *Proceedings of the 22nd international conference on World Wide Web*, 2013: ACM, pp. 1521-1532.

[184] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "An Empirical Study of Geographic User Activity Patterns in Foursquare," *ICwSM,* vol. 11, pp. 70-573, 2011.

[185] Y. Zhuang, S. Fong, M. Yuan, Y. Sung, K. Cho, and R. K. Wong, "Location-based big data analytics for guessing the next Foursquare check-ins," *The Journal of Supercomputing,* pp. 1-16, 2016.

[186] H. Gao, J. Tang, and H. Liu, "gSCorr: modeling geo-social correlations for new check-ins on location-based social networks," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012: ACM, pp. 1582-1586.

[187] I. Zheludev, R. Smith, and T. Aste, "When can social media lead financial markets?," *Scientific reports,* vol. 4, p. 4213, 2014.

[188] S. Y. Yang, S. Y. K. Mo, and A. Liu, "Twitter financial community sentiment and its predictive relationship to stock market movement," *Quantitative Finance,* vol. 15, no. 10, pp. 1637-1656, 2015.

[189] P. Xie, "Predicting Digital Currency Market With Social Data: Implications Of Network Structure And Incentive Hierarchy," Georgia Institute of Technology, 2017.

[190] M. Aljohani, A. Nisbet, and K. Blincoe, "A survey of social media users privacy settings & information disclosure," 2016.

[191] Meaning Cloud. https://www.meaningcloud.com/developer/apis (accessed 29th October 2017, 2017).

[192] L. A. Gottschalk and G. C. Gleser, *The measurement of psychological states through the content analysis of verbal behavior*. Univ of California Press, 1969.

[193] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," 2015.

[194] "Gender workplace statistics at a glance." Australian Government-Workplace Gender Equality Agency. https://www.wgea.gov.au/sites/default/files/Stats_at_a_Glance.pdf (accessed 21st October 2017, 2017).

[195] J. R. Meloy and A. J. Yakeley, "Antisocial personality disorder," *A. A,* vol. 301, no. F60, p. 2, 2011.

[196] P. Liu, J. Guberman, L. Hemphill, and A. Culotta, "Forecasting the presence and intensity of hostility on Instagram using linguistic and social features," *arXiv preprint arXiv:1804.06759,* 2018.

[197] M. Drouin and D. A. Miller, "Why do people record and post illegal material? Excessive social media use, psychological disorder, or both?," *Computers in Human Behavior,* vol. 48, pp. 608-614, 2015.

[198] R. Singh, Y. Zhang, and H. Wang, "Exploring Human Mobility Patterns in Melbourne Using Social Media Data," in *Australasian Database Conference*, 2018: Springer, pp. 328-335.

[199] J. Huang, M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang, "A probabilistic method for emerging topic tracking in microblog stream," *World Wide Web,* vol. 20, no. 2, pp. 325-350, 2017.

[200] A. E. Fahy, S. A. Stansfeld, M. Smuk, N. R. Smith, S. Cummins, and C. Clark, "Longitudinal associations between cyberbullying involvement and adolescent mental health," *Journal of Adolescent Health,* vol. 59, no. 5, pp. 502-509, 2016.

[201] B. W. Fisher, J. H. Gardella, and A. R. Teurbe-Tolon, "Peer cybervictimization among adolescents and the associated internalizing and externalizing problems: a meta-analysis," *Journal of youth and adolescence,* vol. 45, no. 9, pp. 1727-1743, 2016.

[202] K. N. Wang, J. S. Bell, E. Y. Chen, J. F. Gilmartin-Thomas, and J. Ilomäki, "Medications and prescribing patterns as factors associated with hospitalizations from long-term care facilities: a systematic review," *Drugs & aging,* vol. 35, no. 5, pp. 423-457, 2018.

[203] R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," *Computers in human behavior,* vol. 26, no. 3, pp. 277-287, 2010.

[204] R. Didden *et al.*, "Cyberbullying among students with intellectual and developmental disability in special education settings," *Developmental neurorehabilitation,* vol. 12, no. 3, pp. 146-151, 2009.

[205] J. Juvonen and E. F. Gross, "Extending the school grounds?—Bullying experiences in cyberspace," *Journal of School health,* vol. 78, no. 9, pp. 496-505, 2008.

[206] T. Beran and Q. Li, "The relationship between cyberbullying and school bullying," *The Journal of Student Wellbeing,* vol. 1, no. 2, pp. 16-33, 2008.

[207] R. M. Kowalski and S. P. Limber, "Psychological, physical, and academic correlates of cyberbullying and traditional bullying," *Journal of Adolescent Health,* vol. 53, no. 1, pp. S13-S20, 2013.

[208] M. Campbell, B. Spears, P. Slee, D. Butler, and S. Kift, "Victims' perceptions of traditional and cyberbullying, and the psychosocial correlates of their victimisation," *Emotional and Behavioural Difficulties,* vol. 17, no. 3-4, pp. 389-401, 2012.

[209] E. Aboujaoude, M. W. Savage, V. Starcevic, and W. O. Salame, "Cyberbullying: Review of an old problem gone viral," *Journal of adolescent health,* vol. 57, no. 1, pp. 10-18, 2015.

[210] K. Suzuki, R. Asaga, A. Sourander, C. W. Hoven, and D. Mandell, "Cyberbullying and adolescent mental health," *International journal of adolescent medicine and health,* vol. 24, no. 1, pp. 27-35, 2012.

[211] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychological bulletin,* vol. 140, no. 4, p. 1073, 2014.

[212] C. Hay and R. Meldrum, "Bullying victimization and adolescent self-harm: Testing hypotheses from general strain theory," *Journal of youth and adolescence,* vol. 39, no. 5, pp. 446-459, 2010.

[213] K. Hawton, K. Rodham, and E. Evans, *By their own young hand: Deliberate self-harm and suicidal ideas in adolescents*. Jessica Kingsley Publishers, 2006.

[214] M. Vajani, J. L. Annest, A. E. Crosby, J. D. Alexander, and L. M. Millet, "Nonfatal and fatal self-harm injuries among children aged 10–14 years—United States and Oregon, 2001–2003," *Suicide and Life-Threatening Behavior,* vol. 37, no. 5, pp. 493-506, 2007.

[215] E. K. Englander and A. M. Muldowney, "Just Turn the Darn Thing Off: Understanding Cyberbullying," in *Proceedings of persistently safe schools: The 2007 national conference on safe schools*, 2007.

[216] D. C. Kerr, L. D. Owen, K. C. Pears, and D. M. Capaldi, "Prevalence of suicidal ideation among boys and men assessed annually from ages 9 to 29 years," *Suicide and Life-Threatening Behavior,* vol. 38, no. 4, pp. 390-402, 2008.

[217] R. Singh *et al.*, "A Framework for Early Detection of Antisocial Behavior on Twitter Using Natural Language Processing," in *Conference on Complex, Intelligent, and Software Intensive Systems*, 2019: Springer, pp. 484-495.

[218] F. Chollet, "Keras," ed, 2015.

[219] D. T. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[220] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 759-760.

[221] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint arXiv:1412.1058,* 2014.

[222] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1017-1024.

[223] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259,* 2014.

[224] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *2013 IEEE workshop on automatic speech recognition and understanding*, 2013: IEEE, pp. 273-278.

[225] S. Undavia, A. Meyers, and J. E. Ortega, "A comparative study of classifying legal documents with neural networks," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2018: IEEE, pp. 515-522.

[226] A. Agrawal, "Clickbait detection using deep learning," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016: IEEE, pp. 268-272.

[227] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," in *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, 2017: IEEE, pp. 364-371.

[228] J. Seering, R. Kraut, and L. Dabbish, "Shaping pro and anti-social behavior on twitch through moderation and example-setting," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017: ACM, pp. 111-125.

[229] P. Liu, J. Guberman, L. Hemphill, and A. Culotta, "Forecasting the presence and intensity of hostility on Instagram using linguistic and social features," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[230] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proceedings of the First*

Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 1-11.

[231] F. K. Ventirozos, I. Varlamis, and G. Tsatsaronis, "Detecting aggressive behavior in discussion threads using text mining," in *International Conference on Computational Linguistics and Intelligent Text Processing*, 2017: Springer, pp. 420-431.

[232] B. Cao and W.-Y. Lin, "How do victims react to cyberbullying on social networking sites? The influence of previous cyberbullying victimization experiences," *Computers in Human Behavior,* vol. 52, pp. 458-465, 2015.

[233] C. Van Hee *et al.*, "Automatic detection of cyberbullying in social media text," *PloS one,* vol. 13, no. 10, p. e0203794, 2018.

[234] N. Tahmasbi and A. Fuchsberger, "Challenges and Future Directions of Automated Cyberbullying Detection," 2018.

[235] E. V. Altay and B. Alatas, "Detection of Cyberbullying in Social Networks Using Machine Learning Methods," in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, 2018: IEEE, pp. 87-91.

[236] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," *arXiv preprint arXiv:1503.03909,* 2015.

[237] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection with participant-vocabulary consistency," *Social Network Analysis and Mining,* vol. 8, no. 1, p. 38, 2018.

[238] S. T. Aroyehun and A. Gelbukh, "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 2018, pp. 90-97.

[239] N. Craker and E. March, "The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours," *Personality and Individual Differences,* vol. 102, pp. 79-84, 2016.

[240] J. de-la-Pena-Sordo, I. Santos, I. Pastor-López, and P. G. Bringas, "Filtering Trolling Comments through Collective Classification," in *International Conference on Network and System Security*, 2013: Springer, pp. 707-713.

[241] S. Subramani, H. Wang, H. Q. Vu, and G. Li, "Domestic violence crisis identification from facebook posts based on deep learning," *IEEE access,* vol. 6, pp. 54075-54085, 2018.

[242] M. A. Cameron, R. Power, B. Robinson, and J. Yin, "Emergency situation awareness from twitter for crisis management," in *Proceedings of the 21st International Conference on World Wide Web*, 2012: ACM, pp. 695-698.

[243] E. March, R. Grieve, J. Marrington, and P. K. Jonason, "Trolling on Tinder®(and other dating apps): Examining the role of the Dark Tetrad and impulsivity," *Personality and Individual Differences,* vol. 110, pp. 139-143, 2017.

[244] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 1217-1230.

[245] M. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2013: IEEE, pp. 1-5.

[246] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine learning and applications and workshops*, 2011, vol. 2: IEEE, pp. 241-244.

[247] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *European Conference on Information Retrieval*, 2013: Springer, pp. 693-696.

[248] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management,* vol. 24, no. 5, pp. 513-523, 1988.

[249] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* vol. 18, no. 11, pp. 613-620, 1975.

[250] M. Peng, Q. Xie, H. Wang, Y. Zhang, and G. Tian, "Bayesian sparse topical coding," *IEEE Transactions on Knowledge and Data Engineering,* vol. 31, no. 6, pp. 1080-1093, 2018.

[251] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research,* vol. 50, pp. 723-762, 2014.

[252] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, 2012: Association for Computational Linguistics, pp. 90-94.

[253] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency," *Frequency and the emergence of linguistic structure,* vol. 45, p. 137, 2001.

[254] Z.-G. Chen, Z.-H. Zhan, H. Wang, and J. Zhang, "Distributed individuals for multiple peaks: A novel differential evolution for multimodal optimization problems," *IEEE Transactions on Evolutionary Computation,* 2019.

[255] Y.-H. Zhang, Y.-J. Gong, Y. Gao, H. Wang, and J. Zhang, "Parameter-Free Voronoi Neighborhood for Evolutionary Multimodal Optimization," *IEEE Transactions on Evolutionary Computation,* vol. 24, no. 2, pp. 335-349, 2019.

[256] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882,* 2014.

[257] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188,* 2014.

[258] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 workshop on machine learning for information filtering*, 1999, vol. 1, no. 1: Stockholom, Sweden, pp. 61-67.

[259] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *arXiv preprint arXiv:1605.05101,* 2016.

[260] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850,* 2013.

[261] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *International Conference on Artificial Neural Networks*, 2005: Springer, pp. 799-804.

[262] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 85-90.

[263] N. Pogrebnyakov and E. Maldonado, "Identifying emergency stages in Facebook posts of police departments with convolutional and recurrent

neural networks and support vector machines," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017: IEEE, pp. 4343-4352.

[264] S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang, and H. Shakeel, "Deep learning for multi-class identification from domestic violence online posts," *IEEE Access,* vol. 7, pp. 46210-46224, 2019.

[265] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *arXiv preprint arXiv:1702.01923,* 2017.

[266] J. Trofimovich, "Comparison of neural network architectures for sentiment analysis of russian tweets," in *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue*, 2016, pp. 50-59.

[267] G. Gkotsis *et al.*, "Characterisation of mental health conditions in social media using Informed Deep Learning," *Scientific reports,* vol. 7, p. 45141, 2017.

[268] J. Risch and R. Krestel, "Aggression identification using deep learning and data augmentation," in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 2018, pp. 150-158.

[269] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European Conference on Information Retrieval*, 2018: Springer, pp. 141-153.

[270] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673,* 2019.

[271] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[272] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[273] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759,* 2016.

[274] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research,* vol. 11, no. Jul, pp. 2079-2107, 2010.

[275] N. Reimers and I. Gurevych, "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks," *arXiv preprint arXiv:1707.06799,* 2017.

[276] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265-283.

[277] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in neural information processing systems*, 2016, pp. 1019-1027.

[278] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002: Association for Computational Linguistics, pp. 79-86.

[279] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging,* vol. 35, no. 5, pp. 1285-1298, 2016.