EXTRACTING ACTIONABLE KNOWLEDGE FROM DOMESTIC VIOLENCE DISCOURSE ON SOCIAL MEDIA

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

College of Engineering and Science

Victoria University

by

Sudha Subramani

March 2019

ABSTRACT

EXTRACTING ACTIONABLE KNOWLEDGE FROM DOMESTIC VIOLENCE DISCOURSE ON SOCIAL MEDIA

Sudha Subramani, Ph.D.

Victoria University 2019

Respect for human rights is the cornerstone of strong communities, based on the principles of dignity, equality, and recognition of inherent value of each individual. Domestic Violence, ranging from physical abuse to emotional manipulation, is worldwide considered as the violation of the elementary rights to which all human beings are entitled to. As one might expect, the consequences for its victims are often severe, far-reaching, and long-lasting, causing major health, welfare, and economic burden. Domestic Violence is also one of the most prevailing forms of violence, and due to the social stigma surrounding the issue particularly challenging to address.

With the emergence and expansion of Social Media, the substantial shift in the support-seeking and the support-provision pattern has been observed. The initial barriers in approaching healthcare professionals, i.e. personal reservations, or safety concerns, have been effectively addressed by virtual environments. Social Media platforms have quickly become crucial networking hubs for violence survivors as well as at-risk individuals to share experiences, raise concerns, offer advice, or express sympathy. As a result, the specialized support services groups have been established with the aim of pro-active reach-out to potential victims in time-critical situations. Given the high-volume, high-velocity and high-variety of Social Media data, the manual posts evaluation has

not only become inefficient, but also unfeasible in the long-term. The conventional automated approaches reliant on pre-defined lexicons, and hand-crafted feature engineering proved limited in their classification performance capability when exposed to the challenging nature of Social Media discourse. At the same time, Deep Learning the state-of-the-art sub-field of Machine Learning has shown remarkable results on text classification tasks. Given its relative recency and algorithmical complexity, the implementation of Deep Learning-based models has been vastly under-utilised in practical applications. In particular, no prior work has addressed the problem of fine-grained user-generated content classification with Deep Learning in Domestic Violence domain.

The study introduces novel 3-part framework aimed at *(i) binary detection of critical situations; (ii) multi-class content categorization; and (ii) Abuse Types and Health Issues extraction from Social Media discourse.* The classification performance of state-of-the-art models is improved through the domain-specific word embeddings development, capable of precise relationships between the words recognition. The prevalent patterns of abuse, and the associated health conditions are efficiently extracted to shed the light on violence scale and severity from directly affected individuals. The approach proposed marks a step forward towards effective prevention and mitigation of violence within the society.

DOCTOR OF PHILOSOPHY DECLARATION

I, Sudha Subramani, declare that the PhD thesis entitled *Extracting Actionable Knowledge from Domestic Violence Discourse on Social Media* is no more than 100,000 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 22/03/2019

This thesis is dedicated to my parents for their love, endless support and encouragement and my beloved baby girl.

ACKNOWLEDGEMENTS

I owe my gratitude to all those, who have made this thesis possible. First and foremost, I am sincerely grateful to Professor Hua Wang for having been the best supervisor I could have possibly asked for. Thank you for believing in me, and for accepting me as your doctoral student. Thank you for sharing your knowledge, continuous support, kindness, understanding, and patience. Thank you for allowing me to explore my ideas, and choosing my specific research topic that I had desired to pursue. Also, thank you for giving me the support for the internship in Japan. Without your support, I would not have been able to write and submit my research papers, and I could not have attended the conferences, and published in two Q1 journals. Thank you for sharing many important professional, and personal advice throughout these years, and for being there for me in sadness and happiness, and in failure and success. A big thank you from the bottom of my heart for making me what I am today.

I am particularly grateful to Professor Yanchun Zhang, Director of Centre for Applied Informatics, for your valuable advice, encouragement, and support during the team meetings. I would also like to extend my thanks to Professor Yuan Miao, Head of Information Technology for the encouragement and support. Thank you for the motivation to present my research during VU Open Day event. Thank you so much Professors, as the guidance from both of you was invaluable to me.

Thank you Dr. Manjula O'Connor from University of Melbourne and Director of Australasian Centre for Human Rights and Health for sharing the knowledge, suggesting ideas and offering guidance in the Domestic Violence domain. You have inspired me, since the first time I heard about you. Your role as social activist, the awareness campaignes about the problem of violence, education,

and advocacy for human rights, in particular women in abusive relationships, have considerably shaped the direction of my research.

I am thankful to Dr. Huy Quan Vu for your constructive, and encouraging feedback - my first journal paper in particular. Thank you Dr. Rui Zhou, Dr. Siuly, Dr. Huai Liu, and Dr. Khandakar Ahmed for your insightful thoughts that significantly contributed to my research. Thank you Librarian Cameron for offering me the Research Ambassador position at the University, which I am always delighted to work in. Sincere thanks to Dr. Sridhar and Dr. Suganya for your initial guidance regarding the PhD program admission into the Australian university, as well as my candidature proposal.

Thank you my bestie Sandra, who has brought the smile on my face the moment I mentioned her name. My PhD friend, or rather my soul sister (though, you are from Poland), who has always been loving, supportive, and the best part of my university life. Thanks for cheering me up in the stressful time, overnight company in labs, and your tiredless proofreading and recursive writing to every revise and resubmit comment during our paper submissions. Always got the inspiration from your yoga talks and food tips, though I never followed it. Thanks for the support given by my best friend Sarathkumar Rangarajan, whom I have known for nearly 9 years - our journey continues from the master degree to PhD at the same universities. Thanks to my other PhD friends Nithya, Haroon, Alice, Ruwangi and Kevin for their support.

The research presented in this thesis was supported by International Postgraduate Research Scholarship, offered by Victoria University. This financial support is gratefully acknowledged. I want to express my gratitude to Research Dean Anne-Marie Hede, Elizabeth Smith and Dr. Lesley Birch for their constant support and proper guidance.

I am forever grateful to my family friends, dear Venkatesh uncle, Thiagarajan brother and Akila sister for your unconditional love and your support at the right time have brought me this far. A profound gratitude to Thiagarajan brother in particular, for guiding me to prepare my candidature proposal in Data Science for my PhD admission, just a few years back.

Last, but not least, I thank my family for their unconditional love and generous support. Thank you my beloved brother Vivek for your love from afar. Though you do not make any regular calls like our parents, we communicate somehow telepathically. Thanks dear Ezhil for giving me some talking theraphies, whenever I needed to rejunevate myself from stressful times. Thanks Raj, my dear husband, for your support as well.

I am always grateful to my parents, who have sacrificed their life for me and believe the struggle pays off in the future. Thank you mom for all the effort and dedication to raise my child in India, from 6 months to 4 years of age, i.e for the entire duration of my research. Thank you dad for your love and support during the difficult times of my life and believing in me, more than I did in myself. And finally, thank you my lovely baby Naksatra for showering me with cute cuddles and love. Though I could not do any of the motherly duties, and missed all your childhood funs by staying apart. However, during the last few months of this thesis writing, you stayed in Melbourne, and those moments of regular visits to univeristy lab, sitting next to me, and interuppting my keyboard typing with your tiny cute fingers are unforgettable.

PUBLICATIONS

Based on this research work, the following articles, have been published or submitted in International Journals and Conferences.

Journal Articles

1. **Sudha Subramani**, Sandra Michalska, Hua Wang, Yanchun Zhang, Haroon Shakeel, Jiahua Du. "Deep Learning for Multi-Class Identification from Domestic Violence Online Posts". (Accepted: IEEE Access, Mar 7, 2019)
2. **Sudha Subramani**, Hua Wang, Huy Quan Vu, and Gang Li. "Domestic Violence Crisis Identification From Facebook Posts Based on Deep Learning." IEEE access, vol 6 (2018), pp. 54075-54085.
3. **Sudha Subramani** and Manjula O'Connor. Extracting actionable knowledge from domestic violence discourses on Social Media. EAI Endorsed Transactions on Scalable Information Systems, vol 5 (2018), pp. 1-10.

International Conference Articles

1. **Sudha Subramani**, Hua Wang, Md Rafiqul Islam, Anwaar Ulhaq, and Manjula OConnor. "Child Abuse and Domestic Abuse: Content and Feature Analysis from Social Media Disclosures." In 29th Australasian Database Conference, pp. 174-185. Springer, Cham, 2018.
2. **Sudha Subramani**, Sandra Michalska, Hua Wang, Frank Whittaker, and Benjamin Heyward. "Text Mining and Real-Time Analytics of Twitter Data: A Case Study of Australian Hay Fever Prediction." In 7th International Conference on Health Information Science, pp. 134-145. Springer, Cham, 2018.

3. **Sudha Subramani**, Huy Quan Vu, and Hua Wang. "Intent Classification using Feature Sets for Domestic Violence Discourse on Social Media." In 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 129-136. IEEE, 2017.
4. **Sudha Subramani**, Hua Wang, Sathiyabhama Balasubramaniam, Rui Zhou, Jiangang Ma, Yanchun Zhang, Frank Whittaker, Yueai Zhao, and Sarathkumar Rangarajan. "Mining actionable knowledge using reordering based diversified actionable decision trees." In 17th International Conference on Web Information Systems Engineering, pp. 553-560. Springer, Cham, 2016.

Other Presentations

1. Research Presentation during VU Open Day Program 2018.
2. Finalist of VU '3 Minute Thesis' 2018 Competition.
3. Research Presentation during VU Discover Week 2019.

TABLE OF CONTENTS

Doctor of Philosophy Declaration 5
Acknowledgements 6
Publications iii
Table of Contents v
List of Tables viii
List of Figures ix

1 Introduction **1**
1.1 Background and Motivation 1
1.2 Research Problems 4
1.3 Hypotheses 7
1.4 Contributions and Significance 8
1.5 Thesis Outline 12

2 Background **14**
2.1 Prevalence and Types of Domestic Violence 14
2.2 Domestic Violence Impacts 16
2.2.1 Woman's Health Impact 16
2.2.2 Impact on Children 18
2.2.3 Economic Impact 18
2.3 Strategies towards Domestic Violence Prevention and Mitigation 19
2.3.1 Instant Support Services 19
2.3.2 Healthcare Professionals Intervention 20
2.3.3 Campaigns and Awareness Promotion 20
2.3.4 Barriers in Support Services Seeking 21
2.4 Pain and Power: Complex Relationship between Domestic Violence and Social Media 22
2.4.1 Self-Disclosure and Anonymity 23
2.4.2 Social Support in Online Settings 25
2.4.3 Social Media Viral Campaigns 27
2.5 Violence Impact Estimation 29
2.5.1 Data Types 29
2.5.2 Data Sources 30
2.5.3 Data Collection Limitations 31
2.6 Real-time Analytics of Social Media Data 32
2.7 Chapter Summary 35

3 Computational Techniques Overview **37**
3.1 Text Mining for Social Media Application 37
3.2 Pre-processing Steps in Text Mining 39
3.2.1 Feature Extraction 39
3.2.2 Feature Selection 41

3.3 Text Mining Using Classification Techniques . . . 43
3.4 Machine Learning Algorithms Overview . . . 44
3.4.1 Naive Bayes . . . 44
3.4.2 Decision Tree . . . 45
3.4.3 Support Vector Machines . . . 46
3.4.4 K-Nearest Neighbours . . . 46
3.4.5 Logistic Regression . . . 47
3.5 Deep Learning Algorithms Overview . . . 48
3.5.1 Convolutional Neural Networks . . . 49
3.5.2 Recurrent Neural Networks . . . 51
3.5.3 Long Short-Term Memory networks . . . 52
3.5.4 Gated Recurrent Units . . . 53
3.5.5 Bidirectional LSTMs . . . 54
3.6 Application of Classification Techniques in Social Media Data Analysis . . . 54
3.7 Chapter Summary . . . 57

4 Domestic Violence Crisis Identification from Social Media Posts based on Deep Learning 58
4.1 Introduction . . . 58
4.2 Literature Survey . . . 61
4.3 Methodology . . . 67
4.3.1 Data Extraction . . . 67
4.3.2 Data Labeling . . . 68
4.3.3 Feature Extraction . . . 69
4.3.4 Model Construction . . . 70
4.3.5 Performance Evaluation . . . 72
4.4 Experiment and Analysis . . . 73
4.4.1 Experiment Design . . . 73
4.4.2 Psycholinguistic Features Analysis . . . 76
4.4.3 Accuracy Evaluation . . . 81
4.4.4 Hyper-parameters Evaluation . . . 83
4.4.5 Semantic Coherence Analysis . . . 87
4.5 Summary of Findings . . . 90

5 Deep Learning for Multi-Class Identification from Domestic Violence Online Posts 92
5.1 Introduction . . . 93
5.2 Literature Survey . . . 97
5.3 Methodology . . . 101
5.3.1 Data Extraction . . . 102
5.3.2 Gold Standard Construction . . . 102
5.3.3 Feature Extraction . . . 104
5.3.4 Model Development . . . 106

5.3.5 Performance Evaluation . . . 107
5.4 Experiment Design and Analysis . . . 108
5.4.1 Descriptive Statistics . . . 109
5.4.2 Model Training . . . 112
5.4.3 Accuracy Comparison . . . 113
5.4.4 Hyper-parameters Evaluation . . . 116
5.4.5 Models Visualisations . . . 118
5.4.6 Domain-Specific Embeddings Analysis . . . 123
5.5 Summary of Findings . . . 125

6 Automatic Identification of Abuse Types and Health Issues from Online Domestic Violence Posts using Deep Learning 129
6.1 Introduction . . . 130
6.2 Mining Web for Health-related Knowledge Extraction . . . 134
6.3 Benchmark Dataset Construction . . . 136
6.3.1 Data Extraction . . . 136
6.3.2 Gold Standard Labels Construction . . . 139
6.4 Model Construction . . . 141
6.4.1 Feature Extraction . . . 142
6.4.2 Model Specification . . . 145
6.5 Experimental Design and Results Analysis . . . 146
6.5.1 Model Training . . . 147
6.5.2 Accuracy Evaluation . . . 148
6.5.3 Multi-corpus Training and Prediction Results . . . 150
6.5.4 ATHI Recognition and Error Analysis . . . 152
6.5.5 Domain-specific Embeddings Analysis . . . 152
6.6 Summary of Findings . . . 157

7 Conclusion and Future Work 161
7.1 Summary of Contributions . . . 161
7.2 Study Limitations . . . 168
7.3 Future Research Directions . . . 170

Bibliography 173

LIST OF TABLES

2.1 Health Consequences of DV *(Source: WHO Report [116])* 17
2.2 Mission Statements of Various Online DVCS Groups 23

4.1 Examples of DV Posts and the Corresponding Intent Labels . . . 59
4.2 LIWC Features and the Sample Words used in the Dataset 79
4.3 Mean Scores of Psycholinguistic Features for 2 Classes 79
4.4 Performance Evaluation of LIWC Features 81
4.5 Accuracy of Machine Learning Classifiers with Different Preprocessing Steps 82
4.6 Evaluation Metrics of Classification Models 83
4.7 Examples of DV posts and the Corresponding Predicted Labels . 84
4.8 Accuracy of GRUs and LSTMs with Different Parameters Settings 86
4.9 Words have Significant Difference of Occurrence Likelihood between Classes. 87

5.1 Examples of DV posts and the Corresponding Information Category 94
5.2 Example Systems in the Context of Disaster Response and Hate Speech Detection on Social Media 101
5.3 Exploratory Data Analysis of Multiple Classes 112
5.4 Evaluation Metrics of Classification Models 115
5.5 Accuracy of GRUs and BLSTMs with Different Parameters Settings 117
5.6 Accuracy of Deep Learning Models with GloVe and DV Embeddings of 50 & 300 Dimensions. 121
5.7 Sample Posts of Test Dataset and the Corresponding Prediction Results 122
5.8 Example Words and their Top Similar Words from User Posts using DV and GloVe Embeddings 123

6.1 Descriptive Statistics of 3 Data Sources 136
6.2 Distribution of Annotated Instances in Each Corpus 140
6.3 Examples of DV posts and the Corresponding ATHI Mentions . 141
6.4 Most Frequent Words associated with ATHI Mentions 142
6.5 Performance Evaluation of Deep Learning Models with Pretrained Embeddings 149
6.6 Paired Performance Evaluation for ATHI Mentions over the Three Corpora 151
6.7 Examples DV Posts with the Predicted Results and Error Analysis 153
6.8 Performance Evaluation of Stacked Residual BLSTMs with Domain-Specific Embeddings 153
6.9 Example Words and their Top 15 Similar Words from User Posts using DV and GloVe Embeddings 154

LIST OF FIGURES

1.1 Overall Research Framework 9

2.1 Reddit Policy regarding the Anonymous Submissions (Source: subreddit/abusiverelationships [2]) 25
2.2 Influence of Emotional Support received through Social Media . 27
2.3 Revolutionary Hashtags on Social Media 29

3.1 Convolutional Neural Network on a Sample Post: "I am suffering with abusive relationship and desperately need help" 49

4.1 Architecture of our Proposed Methodology for Intent Classification of Critical and Uncritical Posts using Deep Learning Model . 67
4.2 Parallel Coordinates Plot for Critical and Uncritical Posts 80
4.3 Accuracy of Deep Learning Models at Different Epoches 85
4.4 Correlation of sample words 90

5.1 Architecture of Proposed Approach for Multi-class Identification 101
5.2 Accuracy of Deep Learning Models at Different Epochs. 116
5.3 Visualization of Various Information Categories of DV Dataset using t-SNE w.r.t. GloVe Embedding. *(0-Awareness; 1-Empathy; 2-Fund Raising; 3-General; 4-Personal Story)* 118
5.4 Confusion Matrix of Deep Learning Models w.r.t. GloVe Embedding. *(aw-Awareness; em-Empathy; fr-Fund Raising; gen-General; ps-Personal Story)* 119

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Domestic Violence (DV) refers to the various acts of abuse such as physical, sexual, emotional or any controlling behaviour within an intimate relationship [162]. As per the global estimates of World Health Organisation(WHO), nearly 1 in 3 (35%) of women have experienced some type of violence at least at one point in their lifetime [164].

While DV tends to have a greater prevalence in low-income and non-western countries, it is still endemic in developed countries like Australia. This has a disturbingly higher rate of violence against women, such that nearly 49.5% percentage of women have experienced violence, since the age of 15 [47]. Violence against women and their children causes a serious and durable impact on women and children's health and well-being, and on society as a whole [1]. Apart from the mental and physical damage inflicted on the victims, the estimated economic costs for the combined health, social welfare and administration costs amounted to $21.7 billion a year, in Australia. The projections also suggested that costs can be accumulated to $323.4 billion over a thirty year period from 2014-15 to 2044-45 if no additional step is taken to prevent violence against women [47].

Unsurprisingly, DV has become an overwhelming global burden and led to a series of preventative and mitigative strategies initiated by WHO [162]. These include: *(i) Media and advocacy campaigns to raise awareness and facilitate*

the socio-economic empowerment of women, and (ii) Domestic Violence Crisis Support (DVCS) groups' foundation for early intervention services to at-risk families. Over time, the DVCS groups have become the important hubs for safe disclosures about the stigmatized topics, allowing victims to share their experiences in non-judgmental and supportive environments.

Effective solutions to combat violence in family settings require accurate estimates about the problem and its nature. The traditional survey-based approaches are costly and time-consuming. They also involve only a limited number of participants [21], frequently due to concerns for the safety of the victims. As a result, the official statistics tend to underestimate the scale of the problem, significantly impairing the potential risk factors leading to abusive acts identification.

At the same time, the advent of Social Media platforms has grown into importance as a modern outlet for the victims to share their stories, express the sympathy for each other, or seek the emotional/financial assistance [42]. Needless to say, such knowledge could serve as an invaluable source of information for health-care practitioners and policy makers to facilitate the design of the effective preventative measures, and contribute towards essential societal change and well-being.

Despite the vast amount of source data availability, there is still a shortage of techniques that would enable to automatically extract the valuable information from high volume and velocity textual streams. Manual handling approach fails due to the scalability issue, severely impacting the effectiveness of DVCS groups in timely support provision. Additionally, the content posted is often implicit in nature, expressed in non-formal language, including abbreviations,

misspellings, and multiple variations of the same concept. Conventional, often lexicon-based approaches and hand-crafted feature extraction methods prove effective on direct and explicit expressions in well-structured datasets (e.g. medical records), but oftentimes fail to correctly capture the more indirect and ambiguous forms, highly prevalent in online conversations. Consider two following posts: P_1*: "I just recently got out of a DV relationship."*, *and* P_2*: "I desperately need help. Please read my abusive story and consider helping me, thank you. He physically assaulted me, and threatened to kill me. I have spent the last 10 months with depression, insomnia, and Post Traumatic Stress Disorder (PTSD)."* While P_1 openly indicates the post-abusive experience, P_2 only implies the exposure to the emotional and controlling behavior, potentially leading to further escalation if appropriate measures are not undertaken. Such linguistic variations significantly impair the correct classification and relevant information extraction.

To address the problem stated, advanced Machine Learning techniques are required to accurately identify critical situations from the deluge of incoming data. Process automation allows reducing the response time, minimizing the damaging health impact on the victims, and preventing further harm from happening. The actionable knowledge extraction from Social Media has already proved significant for public health mining [34, 38, 156, 212] and crises scenarios [30, 99, 236], where the information available in the early hours of an event is critical for the disaster management services.

Deep Learning as an advanced sub-field of Machine Learning have demonstrated promising results in text classification tasks [118] due to its ability to automatically capture latent features in character strings. Most recently, word embeddings [142, 181] as part of the feature extraction step in Deep Learning

has further contributed towards performance improvement by accurately identifying the syntactic and semantic relationships between the terms and phrases. Despite high accuracies obtained, Deep Learning is considered a non-intuitive, highly empirical and task-specific technique. The performance of the models is closely linked to the case study investigated. The DV discourse analysis on Social Media still remains a niche application domain with a relatively few number of studies [213]. Only limited research has been conducted so far to investigate in-depth how the state-of-the-art techniques can be utilized and improved to extract the critical information from user-generated online content. The thesis aims to develop the framework of the automatic classification system to transform the highly unstructured discourse streams into targeted knowledge in support of emergency services responsible for life-saving decisions.

1.2 Research Problems

The primary research question is divided into 3 related sub-questions and each detailed below.

Research Question: How the relevant information from DVCS groups perspective can be identified automatically and accurately from Social Media online discourse?

In the past decade, social networking platforms have become the major channel connecting the support-seeking individuals with support-providing services. Given large volume and unstructured format of Social Media data, the first problem to address is to determine what constitutes critical information for DVCS groups. Therefore, the Sub-Question 1 focuses on information

needs identification for future corpora development and model training. As Social Media poses significant challenges due to its high-volume and high-velocity character, the only solution for effective content management and critical posts identification in a timely manner involves process automation. Consequently, the Sub-Question 2 concentrates around the state-of-the-art methods application, validation and improvement in the DV context. Finally, given the tremendous potential that user-generated content holds in addressing DV issues, the text mining approaches and the derived knowledge, such as Abuse Types and Health Impacts, will be addressed by the Sub-Question 3.

Sub-Research Question** 1**: What are the information needs required by DVCS to timely respond and effectively address the crisis situation on Social Media platforms?

In order for the classification system to effectively serve its purpose, the appropriately annotated benchmark corpora is required. The human resources are limited, and the scale of the problem of violence, as expressed on Social Media, is immense. Also, the time proves a critical factor in at-risk situations. The timely detection and accurate automatic prediction of type of the support needed proves invaluable in addressing the high-velocity data. Given the availability of task-specific annotated corpora significantly reduces the time of model development and implementation. Furthermore, system optimization and validation additionally ensures its practical application on real-world case scenario. Currently, no fine-grained DV benchmark corpora, aligned with the DVCS information needs exists. Thus, the "Gold Standard" dataset will be developed under the guidance of the domain expert as part of this thesis as well as for the future research in the DV field.

Sub-Research Question 2: What Natural Language Processing and Machine Learning approaches yield the highest accuracy on text classification task within DV dataset, and how their performance can be further improved?

After identifying the information needs and the expected outputs of the system, the next question is how the classification task can be performed automatically to achieve human-like accuracy in posts evaluation process. As the manual approach proves not only inefficient, but eventually infeasible given the large-scale streaming sources, the automation step is rendered necessary in continuation of online support services. Furthermore, the specific characteristics of Social Media data prove the task challenging, adversely affecting the performance of "shallow" classifiers with conventional feature extraction methods. For that reason, the state-of-the-art Deep Learning models will be developed and validated against themselves and the traditional Machine Learning algorithms. The most advanced feature extraction methods (i.e. word embeddings) will be implemented for further performance improvement. As the recent studies report the observed accuracy increase following the domain-specific embeddings development, the DV embeddings will be generated and evaluated. The findings will prove invaluable in practical application of the system designed.

Sub-Research Question 3: What knowledge can we derive about the violence problem from large scale Social Media data applying text mining techniques to support healthcare practitioners and policy makers?

Despite the increasing popularity of self-disclosure and support-seeking among DV victims, the limited research exists with regard to actionable knowledge extraction in DV domain. While the information about few individuals does not provide enough insight into the problem of violence, the large-scale

data can reveal hidden trends and patterns prevalent within society. The unstructured format of text data requires advanced natural language processing techniques to ensure that derived knowledge is meaningful and offer real benefit to health-care practitioners and policy makers. Therefore, following the automatic classification process, the Abuse Types and Health Issues will be mined from the relevant posts identified. Both information categories *(i.e. Abuse Types and Health Issues)* have been selected based on the frequent mentions of particular kind of abuse by the victims *(e.g. physical, psychological, sexual)* as well as the potential health conditions before/after the experience *(e.g. depression, anxiety, PTSD)*. The knowledge derived will support the health organisations in accurate estimations of the scale of the problem, as well as facilitate the targeted services provision.

1.3 Hypotheses

In order to address the aforementioned research questions, the following hypotheses are generated:

1. State-of-the-art Deep Learning models are able to accurately (acc > 0.9) identify the pre-specified classes from highly informal Social Media streams related to DV, and outperform the traditional Machine Learning classifiers.
2. Development of domain-specific word embeddings trained on the large scale DV-related dataset can further improve the classification performance and knowledge discovery in comparison with the default word embeddings.

3. Analysis of the extracted DV data using advanced text mining techniques can reveal the specific syntactic and semantic characteristics of the language used by the victims, as well as provide valuable knowledge about the prevalence and severity of the problem of violence within society (e.g. Abuse Types, and Health Issues).

1.4 Contributions and Significance

Social Media has been increasingly used in violence prevention by awareness raising, knowledge sharing, and bringing stories to the public. Despite the increasing popularity of self-disclosure and support seeking among DV victims, the limited research exists regarding the actionable knowledge extraction in DV domain. The study proposes the highly performing (acc > 0.9) and empirically validated Deep Learning-based system aimed at automatic user-generated content categorization in order to improve the efficacy of DVCS groups crises situations identification and timely victims support.

The particular contributions and their significance have been outlined in the following paragraphs. The overall research framework is presented in Figure 1.1.

Critical Posts Detection: The deluge shared on DV related forums frequently leads to critical posts omission, entailing potentially harmful consequences for the individuals in need. The content ranges from awareness promotion campaigns, advice offering, or expression of empathy. Such posts are considered "non-critical" from DVCS groups perspective, as the person is not at-risk and do not request immediate support. In contrary, certain posts are

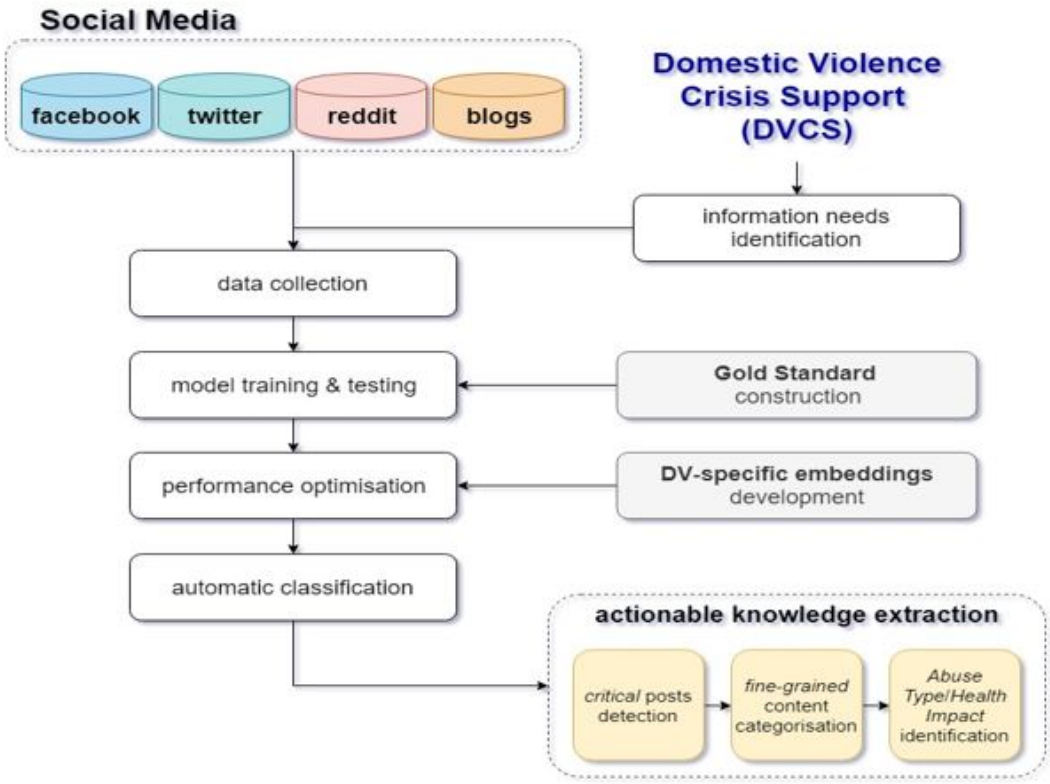


Figure 1.1: Overall Research Framework

treated as "critical", where the victim report the potential of danger and require immediate intervention of support services. The accurate identification of critical posts is crucially important for DVCS to direct their limited resources for support of those in at-risk situations. Manual browsing through a large amount of online content is time consuming, inefficient, and unscalable in longer-term for DVCS moderators. Hence, the automated framework for critical posts identification in DV context has been proposed.

Fine-grained Content Categorization: The automatic content classification allows the DVCS groups to efficiently handle the high volume and high velocity data, evaluate the nature of the problem, and respond almost instantly. Thus, the multi-class posts categorization has been presented, where 5 dis-

tinctive classes are identified, providing the in-depth insight into the violence prevalence and severity. The approach is the extension of the previous work focused on binary posts classification. The finer-grained classes identification enables the online content to be re-routed to appropriate services, improving the efficiency of DVCS groups in addressing abusive instances reported on Social Media.

Abuse Types and Health Issues Identification: As DV victims increasingly share their personal abusive experience on the online DVCS forums, the details of potential violence incidents as well as the related health issues before/after exposure to abuse are also disclosed. Hence, the framework was proposed, that automatically extracts mentions of Abuse Types and Health Issues from victims' posts. Thus, the timely detection of potential DV incidents are indispensably important for DVCS groups to pro-actively reach out the victims in a timely manner.

Gold Standard Datasets Construction: Although benchmark datasets are available in contexts such as mental health or natural disasters, the extensive and high-quality annotated corpora for DV prediction and associated health conditions mining from Social Media is not yet readily available. Hence, the novel gold standard dataset has been constructed under the supervision of the domain expert, active in DV domain. The availability of annotated corpora, designed for both binary and multi-class identification, as well as fine-grained health information extraction forms a basis for effective content categorization. As annotation is considered an expensive and time-consuming process, the release of domain-specific gold standard to public repository allows for instant model training and application. It is also the first annotated corpora for the

tasks specified, as no previous works addressing the issues identified within DV context exists. Furthermore, the gold standard datasets open publication to research communities dedicated to violence prevention and domestic abuse invites further advancements in predictive accuracy improvement.

Domain-specific Embeddings Construction: The novel domain-specific embeddings have been developed to explore their effectiveness in classification accuracy improvement following the promising results reported in previous studies. The DV embeddings have been trained on approx. 500k posts, collected from various online user-generated contents (i.e. Facebook, Twitter, Reddit, blogs etc.). The developed embeddings were evaluated in terms of *(i) impact on the predictive performance of the applied Deep Learning models, (ii) valuable insight generation and knowledge discovery about the problem of DV.*

Actionable Knowledge Extraction: The study has placed an emphasis on actionable knowledge extraction, focusing on the alignment with the DVCS groups information needs in order to improve the efficiency of the support services provided. The automated content categorization into the pre-defined classes allows the DVCS moderators to handle the volume effectively by rapid assessment of the nature of the problem, along with the support needed. The text mining techniques have explored the linguistic characteristics observed in the abusive acts descriptions by the victims. The knowledge about particular types of abuse dominant in the society serves as an invaluable source of information for health care professionals and government officials for targeted preventative initiatives development. The automated extraction of the health conditions associated with the experience of abuse enables the health care professionals to identify the dominant risk factors and warning signs leading to

critical situations, as well as the potential impacts on victims mental and physical condition following the exposure to violence. Such knowledge serves as an invaluable complementary source of information for the official health statistics, leading to the appropriate services provision.

1.5 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 outlines the background about the wider prevalence of DV and it's impacts, the various preventive strategies initiated by WHO, discusses about the power of Social Media in supporting the stigmatized contexts, the various types and sources of data required to measure violence and it's impacts and defines brief about other computational studies that used Social Media data.

Chapter 3 discusses about the various pre-processing steps in text mining tasks, brief explanation on state-ot-the-art Machine Learning and Deep Learning techniques and their applications in text classification tasks on Social Media.

Chapter 4 presents an approach for "critical posts identification" from user-generated online discourse. Firstly, a benchmark dataset of posts extracted from Social Media is constructed by manually labelling the content as either "critical" or "non-critical" for binary classification. Textual features are subsequently extracted from the unstructured textual strings for further processing. Deep Learning models are adopted as the state-of-the-art classification approaches. The performance of the proposed model is evaluated against the various feature extraction approaches as well as other conventional Machine Learning tech-

niques. Furthermore, the most informative and distinctive features between "critical" and "non-critical" posts highlighted for valuable insight into the DV problem.

Chapter 5 introduces a framework for "multi-class automatic posts categorization", providing the finer-grained insights into the specific categories of posts shared on DV dedicated online forums. Firstly, 5 distinctive classes are identified according to DVCS information needs, following by the benchmark corpora construction. The state-of-the-art Deep Learning models are subsequently trained. The accuracies obtained are compared between the 5 Deep Learning architectures as well as with the traditional Machine Learning techniques as a baseline. Additionally, the specifically developed domain-specific embeddings are applied in the feature extraction step during models training to validate their effectiveness in classification performance improvement.

Chapter 6 proposes the system for "automatic identification of particular Abuse Types as well as associated Health Issues" from victims posts shared on online DVCS forums. Firstly, the dataset is narrowed down to the critical posts, following the procedure explained in Chapter 4. Then, the experiments with 2 state-of-the-art Deep Learning architectures are conducted, along with the domain-specific and default word embeddings, as modern feature extraction approaches proved successful in capturing the syntactic and semantic relationships between the terms. Finally, the most prevalent Abuse Types and Health Impacts are identified.

Chapter 7 summaries the findings and concludes the results obtained from the combined framework and indicates the new directions for the future improvements.

CHAPTER 2

BACKGROUND

This chapter outlines the background setting for DV problem, in particular - its prevalence and severity within the society, its leading causes and observed consequences, and finally the preventative and mitigative strategies, initiated by WHO. Furthermore, the power of recently emerged Social Media phenomenon in support of painful and stigmatised issues is discussed. The various types of data and its sources are also presented, in the context of DV impact estimation. Lastly, the examples of studies utilising user-generated content for actionable knowledge extraction in time-critical situations in application to social issues are presented.

2.1 Prevalence and Types of Domestic Violence

DV involves violent, abusive or intimidating behavior by a partner or family member, to control, dominate or cause fear to other family members [162]. DV is one of the most prevailing forms of violence, and has become an overwhelming global burden. The definition also encompasses Intimate Partner Violence (IPV), which forms a sub-group of the cases, where violent acts take place between the intimate partners/spouses specifically. Hence, both the terms, i.e. DV and IPV, are often used interchangeably throughout this thesis.

DV is the most prevalent form of violence experienced by women. Women are also more likely to be abused at home by a husband or male partner than anywhere or by anyone else [116,162,163]. Whereas in the case of men, they are more likely to be attacked by a stranger or acquaintance than by someone within

an intimate relationship. Though the women tend to be abusive in relationships with men or other women (same sex relationships), the most prevailing forms of violence acts were borne by women at the hands of abusive men. This is considered a global issue of pandemic proportions, regardless the age group, religious belief, cultural belonging, socio-economic status or geographic location. The WHO estimates that 35% of women worldwide have experienced abuse, and this has become an issue of global concern [162]. The problem of violence is increasingly elevated as a substantial public health problem due to its devastating consequences, initially considered to be a violation of human rights.

According to previous population-based surveys around the world, between 10% to 69% of women reported physical abuse by intimate male partner at some point in their lifetime [116]. The surveys further state that majority of the victims experience multiple acts of violence, mostly physical abuse that is often accompanied by psychological and/or sexual abuse. The consequences of violence are profound, affecting not only the health and happiness of its victims, but also the well-being of the entire communities. The consequences of the DV are discussed in next section.

DV refers not only to physical acts, but also includes sexual, verbal, psychological and financial abuse [162]. The main types of violent behaviours with the examples are as follows:

- **Physical Abuse:** beating, biting, slapping, hitting, beating etc.;
- **Sexual Abuse:** forced intercourse, rape (marital or attempted rape), sexual harassment or sexual threats;
- **Psychological or Emotional Abuse:** such as constant criticism, manipulation, coercive control, shame and blame;

- **Financial Abuse:** family income control and prohibition to family funds access;
- **Stalking:** Harassment or threatening tactics such as unwanted phone calls, messages or emails, causing fear and safety concerns among the victims.

2.2 Domestic Violence Impacts

Violence against women causes serious and durable impact on women, the children that witness an abuse within family, their health and well-being, and effectively the society as a whole [237]. Recently, the policy and public discussions on violence against women have had a strong focus on DV [129]. Though economic costs are calculable and provide concrete metrics for policy makers, the true costs associated with the well-being (physical or psychological) of the victims can be exacerbated and unpredictable. The health costs of violence as per the Burden of Disease methodology are significant [237]. As an example, IPV contributes to more death, disability, and illness in women aged 15 to 44 than any other preventable risk factor. Also, the women suffer from physical injuries, mental illnesses and reproductive health issues [161,165,237]. The various types of violence impacts on its victims are discussed in the following sub-sections.

2.2.1 Woman's Health Impact

Violence has profound effects on women's health, both immediate and long term. Table 2.1 draws on the scientific literature that summarises the host of

different health issues that are linked with IPV [116, 162, 163]. Although violence has immediate and visible health impact, ranging from severe injuries to minor bruises, being a victim also increases a women's risk of suffering from long term and invisible mental health issues, such as depression, anxiety or an adoption of risk behaviors including alcohol consumption or drug use.

Table 2.1: Health Consequences of DV *(Source: WHO Report [116])*

Physical **[A]**	Mental & Behavioral **[B]**	Reproductive **[C]**	Fatal **[D]**
Injuries	Depression and anxiety	Gynaecological disorders	AIDS-related mortality
Bruises and cuts	Post-traumatic stress disorder	Infertility	Maternal mortality
Chronic pain syndromes	Psychosomatic disorders	Pelvic inflammatory disease	Homicide
Fibromyalgia	Phobias and panic disorder	Pregnancy complications/miscarriage	Suicide
Fractures	Poor self-esteem	Sexual dysfunction	
Gastrointestinal disorders	Suicidal behaviour and self-harm	Sexually transmitted diseases, including HIV/AIDS	
Irritable bowel syndrome	Feelings of shame and guilt	Unsafe abortion	
Lacerations	Alcohol and drug abuse	Unwanted pregnancy	
Ocular damage	Smoking		
Reduced physical functioning	Eating and sleep disorders		
Disability			

- **[A] Physical Health:** According to population based studies, approx. 40−72% of women, who have been physically abused by a partner suffered from cuts and bruises. However, 'injury' is not the most common result of physical abuse. Functional disorders (a host of ailments that occur with no identifiable medical cause) such as fibromyalgia, gastrointestinal disorders, functional dyspepsia and irritable bowel syndrome, and various chronic pain syndromes are more prevalent among women reporting an abusive instance [25].
- **[B] Mental Health:** Beyond immediate repercussion, abused women suffer more from mental health issues such as depression, phobias and anxiety than non-abused women. Women further suffer from psychological well-being and the adoption of risk behaviors, such as smoking, physical inactivity, alcohol and drug use. Also, women abused by their partners have an increased probability of homicide, suicide or suicide attempts [26, 28]

- **[C] Reproductive Health:** Women have difficult time protecting themselves from unwanted sex, when living with abusive partners. Sexual abuse can lead to unwanted pregnancy, an increased risk of HIV infections, and effectively - contributing towards AIDS epidemics. The violence that occurs during pregnancy not only impacts woman's health, but also affects the developing fetus, resulting in miscarriage, stillbirth, premature birth, fetal injury, and low birth weight [81,130,211].

2.2.2 Impact on Children

Children are often present during the incidents of violence and routinely witness the abusive behaviour. The violence at home directly/indirectly affects children's development, placing them at a higher risk of emotional and behavioral issues such as depression, anxiety, nightmares, poor school performance, and low self-esteem [91,92].

2.2.3 Economic Impact

In addition to the severe health consequences of women and their children, violence also have serious impact on victims professional development [46,189]. Their job performance and overall employment probability of is severely diminished. This places an huge economic burden on societies due to both - a loss of productivity and increased utilisation of social support services. Thus, the DV greatly influences women's earnings and personal income, leading to a reliance on the financial assistance from the community or government.

2.3 Strategies towards Domestic Violence Prevention and Mitigation

Increasing victims access to support services is a pressing need around the globe [62]. Hence, WHO has framed the number of initiatives, including: strengthening appropriate support services to victims, and promoting awareness campaigns to achieve global reduction in violence rates [103]. Various support services and community-based efforts are discussed in the following sub-sections.

2.3.1 Instant Support Services

The support services required may vary depending on the circumstances of the victims of DV. During crisis, women tend to seek emotional support (sympathetic encouragement or counseling), financial aid or legal counseling (about divorce, child custody or maintenance options) [140]. They further require temporary shelter due to an increased likelihood of homelessness. The safe haven is crucial for urgent protection, especially during leaving the violent relationship, as it might be a dangerous time for the women. As a result, the various crisis support and shelter centers offer specialised services for the victims that are either supported by government or privately-run organisations. However, it is vital for the centers to ensure that the range of services provided is accessible and coordinated properly to fulfill the needs of victims in distress.

2.3.2 Healthcare Professionals Intervention

As discussed in previous sub-section 2.2.1, violence against women leads to multitudes of negative impacts on physical, mental and reproductive health. Hence, the overwhelming attention has turned towards the healthcare providers to improve their response that includes identifying the victims of violence, providing appropriate care for victims, and referring them to the specialised services, if required [101]. Unfortunately, according to previous studies, healthcare professionals rarely check for obvious signs of abuse, and rarely enquire the patients about the possible histories of violence experience [69]. Hence, those studies highlight the need for sensitising the healthcare providers to improve the identification and response rate to DV victims by drawing up clear information sharing protocols for proper management of abuse instances records.

2.3.3 Campaigns and Awareness Promotion

Campaigns tend to raise awareness about the problem of DV within the wider community, in order to intervene and pro-actively prevent partner or family violence. They also enable the victims to access various range of the support services (health, safety and legal) that are particularly sought-after in crisis situations [27, 128].

The following describes the primary aims of the campaigns that are organised by government and privately-run organisations:

- To 'break the silence on violence against women' - the first, albeit essen-

tial step towards ending violence, i.e. awareness campaignes, increasing visibility of violence;

- To influence the decision makers to provide more public support and to induce essential policy changes;
- To provide encouragement the victims, who hesitate in abuse reporting, to seek appropriate support services.

2.3.4 Barriers in Support Services Seeking

Women experiencing violence require earlier access to the specialised services such as professional healthcare, crisis support, legal guidance, and so on. As violence is considered the socially stigmatised context in numerous cultural and societal environments [117,126], it hinders the help-seeking behavior of battered women. Many women go through great pains by concealing their sufferings and under-report the abuse instances. The reasons for that range from personal fears of being blamed, disbelieved or isolated. In certain cases, it could be attributed to the lack of financial support, unawareness of the available support services, or lack of information about their legal rights [69]. The reliance on the external enabler (either professional, family member or a friend) to facilitate support services is frequently observed.

Despite an increasing levels of consultation for depression and anxiety, the reference from General Practitioners (GP) is required in order to access specialist support. Informal disclosure to family member or a friend leads to specialist help only, if they themselves were victims of DV, or have the relevant knowledge about it. Women face multiple barriers to active support seeking, and often

need external help to access the dedicated services. The prompt response from healthcare and social professionals is also required [249]. The women recount long periods of ambivalence, disclosing the violence acts only after leaving the perpetrator [69, 182].

2.4 Pain and Power: Complex Relationship between Domestic Violence and Social Media

According to statistical report published by Global Web Index, the number of prolific consumers of Social Media reached approx. 3.2 billion users worldwide in year 2018 [32]. This accounts for nearly one third of the current global population, as of 2019. Currently, a myriad of Social Media platforms such as Twitter, Facebook or Instagram exists. Their gaining popularity permeates every aspects of people's lives and relationships. As an example, the Facebook has attracted approx. 2 billion active users [32]. Due to the growing impact of Social Media, it has been heralded as a medium to share not only the opinions and thoughts, but also to solicit support, aid, and rescue purposes during the times of crises [144]. According to previous studies, DV has unique characteristics of isolating its victims from the outside world [117, 182]. In contrast, Social Media has proven its potent role in facilitating the victims to establish connection to the broader community, which is otherwise lacking in their lives. DVCS organisations harness the power of Social Media platforms, and launch their services online to provide the information, support, and resources to the victims, which is otherwise impossible to access outside the digital world. The Table 2.2 defines the mission statements of the DVCS groups that proclaim their commitment to victims

Table 2.2: Mission Statements of Various Online DVCS Groups

Organisation Name	Media	Mission Statement
The Red Heart Campaign [24]	Facebook	Shines a light on violence against women and children. We share stories of violence survival, provide material aid, lobby for change and deliver daily the latest news, information and views on violence against women and children from Australia and around the world.
Break the Silence Against Domestic Violence [229]	Facebook	To educate communities on the dangers of domestic violence, connect victims and survivors, and assist them in the transformation of their lives.
Domestic Violence Support [224]	Facebook	Come here and share your stories. Get support from other people who have been in your situation or a similar.
r/domesticviolence [60]	Reddit	Information and support for victims, survivors, their friends and family. Are you being abused? We have resources for you. Are you are survivor? We want to hear from you. We offer support and resources.
r/abusiverelationships [2]	Reddit	For everyone (male and female) who has ever been in an abusive relationship or is currently in one. This is a place for people to vent, share their stories and offer support to others in similar situations.

of violence. The established groups not only provide support/promote awareness, but also encourage the victims to share their abusive stories, offer a place to vent, and establish the necessary connections.

The following sub-sections discuss in detail the increasing rate of self-disclosure in the form of personal stories shared by the victims on social Media, supported by the anonymous nature online environments, and various support services received from the community in turn.

2.4.1 Self-Disclosure and Anonymity

Although seeking support from various services (such as advocacy and housing) helps in recovery [135], DV survivors frequently experience severe barri-

ers in active reach-out for assitance. Firstly, the drawback of support-seeking is commonly known as 'losing face'. In Goffmans terms, 'face' is the positive self-image that people present in their social interactions, especially when they wish to maintain their social status [82]. Secondly, DV survivors experience lack of trust in others, feelings of shame, embarrassment. They may deny, or fear of ramifications, such as the revenge from the perpetrator or his family members [70]. Due to a coercion or isolation by the abusive partners, women experiencing violence tend to have small networks and weak interconnections, placing severe limits on their disclosure [226]. Other hindrance to the disclosure includes poor past experiences including being disbelieved, blamed or judged (negatively), and effectively invalidating their experience. In many cultures, women are socialised to believe that it is okay to be emotional, discuss their problems, and seek help, as men are believed to handle the problem on their own [131].

Women are sharing their emotions and personal stories of their abusive relationship increasingly on Social Media rather than face to face interactions, possibly due to the above-mentioned barriers. Hence, there is an increased share of reports and stories on abuse within the dedicated groups on Facebook, Reddit or Twitter. The motivation of sharing may be due to the following reasons: *(1) emotional regulation, i.e. managing emotions through social interaction help people move back into emotional equilibrium [258], and (2) anonymity support and respect people's privacy for the sensitive disclosures encouragement.* For example, Figure 2.1 defines the rules of subreddit forum on the 'abusive relationships', where one of the rules states about people's privacy. According to the study by Natalya et al. [13], people felt more satisfied after sharing emotional exerience through Social Media, which is strongly tied to the aspects such as quality of replies, not

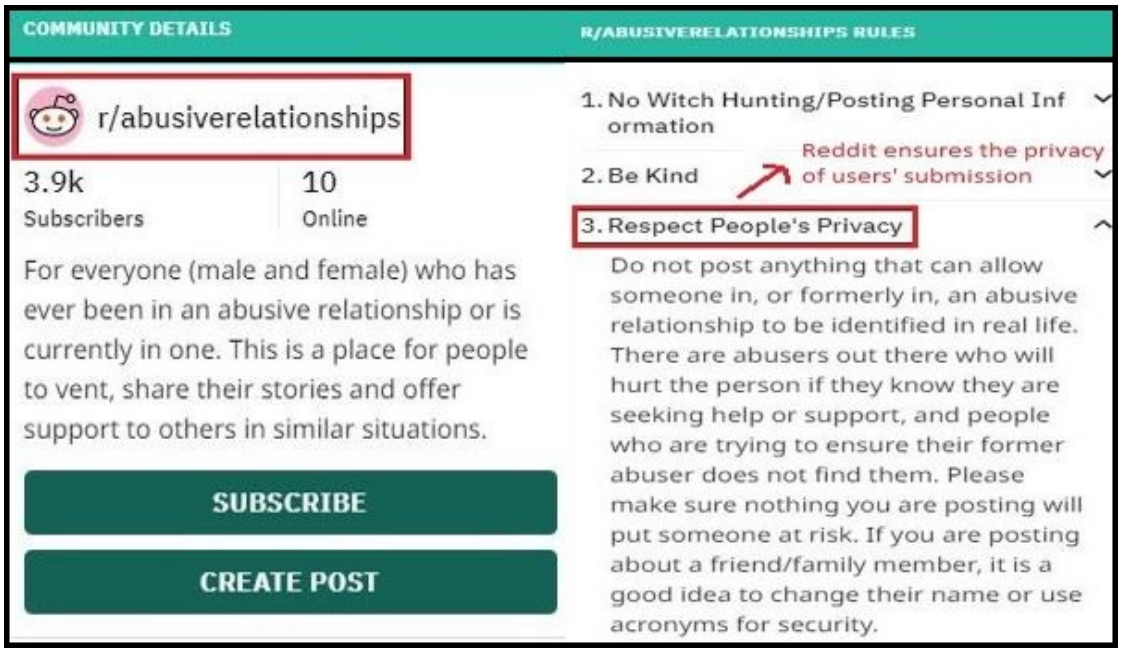


Figure 2.1: Reddit Policy regarding the Anonymous Submissions (Source: subreddit/abusiverelationships [2])

merely of their volume. This is turn suggests the need for the design of supportive environments that help people in need to receive valuable responses to their shared stories.

2.4.2 Social Support in Online Settings

Social Media play a leading role not only in awareness promotion, but also to leverage various dimensions of social support such as emotional, instrumental, and informational. Figure 2.2 depicts the influence of emotional support, received through Social Media. Overwhelming response to the tragic post of lives lost to violence amounted to approx. 5.1k likes, 1k comments, $8,241$. Social support has been demonstrated to trigger numerous positive outcomes including general well-being and life satisfaction [235]. Informational support implies

the exchange of information through the provision of advice, guidance, suggestions, or concrete ways people assist each other [197]. For instance, the support can include the recommendations for an appropriate legal support, healthcare services, or refugee housing. Also, the advice on how to obtain an intervention order during tough times of life can be provided. Sabine et al. [231] observed that Social Media are better suited to exchange informational support than face-to-face communication. Emotional support refers to the expression of positive affect towards people we care, feeling of belonging, and emotional reassurance, sharing of love and empathy. According to Jessica et al. [239], emotional support in online context complements the emotional support received in offline contexts. Facebook users are benefited by conveying support-based needs due to the ease of sharing with a wide variety of connections, and also by efficiently broadcasting information without the effort required involving phone calls or emails [239]. On Facebook, users receive emotional support not only from wide spectrum of friends, but also from people outside their regular support network, benefiting from wider scope of individuals with similar life experiences. As an example, 28 years old woman was active in a Military Wives Group as it allowed her to 'vent' her frustration that even close friends, who did not have military partners, might not have understood [239]. Similarly, Turner et al. [234] finding show that cancer survivor woman finds similar others through the Facebook 'Group' feature. Though the women had no prior relationship with the other women with whom they interacted, they realised that the shared experiences of the group enabled meaningful exchanges, and garnered social support from the other members' comments.

(a) Overwhelming Response Received for Heart-rending Post

(b) Sample of Comments that implies Empathetic Encouragement from the Community

Figure 2.2: Influence of Emotional Support received through Social Media

2.4.3 Social Media Viral Campaigns

The availability of Social Media has challenged the notion of violence as private one [121]. According to Heather et al. [134], Social Media platforms have become the powerful agents for engaging public into a dialogue about the realities of DV. In terms of the self-disclosure, detailed storytelling, direct and indirect support-seeking, and emotional exposure are increasingly observed in virtual environments [5]. As stated by Trepte et al. [231], mental support received in the online contexts visibly complements the support received off-line.

The hashtags *#WhyIStayed* and *#WhyILeft* became trending on Twitter in 2014, where the DV victims shared their stories on why they stayed or left

the abusive relationship [1]. The online posts mining was also successfully employed to identify the factors behind staying/leaving decisions among victims [43, 48, 246]. Another major trend took place in 2016, when Twitter hashtag *#MaybeHeDoesntHitYou* [2] triggered an outpouring of victims stories detailing their personal experience with an abusive behaviour. Following, the *#MeToo* [3] campaign on sexual violence against women went exceptionally viral, with men retweeting the *#HowIWillChange* [4] hashtag in order to shift the perspective regarding the rape culture [184]. With positive momentum initiated by *#MeToo* movement across the globe, in 2018 the UN women and their partners were marking 16 Days of Activism Against Gender-Based Violence. The event was promoted with *#HearMeToo* [5] hashtag on Twitter for promotional purposes. Figure 2.3 demonstrates the examples of how the prominent hashtags inspired the global movement of women to share their stories of DV and sexual assault. Figure 2.3a depicts an example of how the term *#MeToo* gained momentum after actress Alyssa Milano asked victims of sexual assault to come forward and share their experience with sexual assault. The hashtag was posted more than 6 million times on Social Media platforms including Facebook and Twitter between October to December 2017 [6]. Similarly, Figure 2.3b presents the example of hashtag *#WhyIStayed*, which prompted the victims to share their own stories of violence, and it has been used more than 100, 000 times in only 2 days according to Social Media analysis website Topsy [7]. Thus, Social Media campaigns raised against violence play significant role in shaping the openness culture and breaking the silence around the most pressing community issues.

[1] https://twitter.com/hashtag/whyistayed
[2] https://twitter.com/hashtag/maybehedoesnthityou
[3] https://twitter.com/hashtag/metoo
[4] https://twitter.com/hashtag/howiwillchange
[5] https://twitter.com/hashtag/hearmetoo
[6] https://www.bbc.com/news/world-42251490
[7] https://vulcanpost.com/20262/why-i-stayed-why-i-left/

(a) Hashtag #MeToo (b) Hashtag #WhyIStay

Figure 2.3: Revolutionary Hashtags on Social Media

2.5 Violence Impact Estimation

This section describes the various type of data and its sources to measure the impact of violence, along with the constraints of the existing data collection approaches.

2.5.1 Data Types

According to WHO, various types of information are vital for understanding the circumstances surrounding the violent incidents. The types of data include:

- Mortality data about fatalities through homicide and suicide can provide an information on the impact and the extent of toxic relationship within a particular community. This knowledge can also be used to measure the burden created by fatal violence, how it evolves over time, and in the high-risk communities identification.

Given that non-fatal outcomes are more common than fatal outcomes, and

the actual violence rates are not be fully represented by the mortality data, other type of information is necessary that includes:

- Health-related data on diseases, injuries, or other health conditions that describes the impact of violence on health of individuals and groups;
- Self-reported data on victimisation, exposure to violence, attitudes, beliefs, and cultural practices;
- Other data types include community data, crime data, economic data, and outline the population characteristics, including the income level, education, violent events, offenders, as well as economic costs of health and social services.

2.5.2 Data Sources

The different types of information are collected from various data sources. For instance, data from the police departments include the victim-perpetrator relationship, whether a weapon was used and other related circumstances. Whereas, the data from emergency health departments include the nature of injury and how the accident happened. However, statistical surveys are considered to be potential source for the detailed information about the violence act, health-related injuries, attitudes, behaviours, and other possible incidents involved in violence. Thus, the survey studies help to uncover the vital information that is not even reported to police or health departments.

2.5.3 Data Collection Limitations

The constraints observed relate majorly to the scarcity and incompleteness of information available, and can be listed as follows:

Availability and Quality of Data

As different agencies maintain records for their own record-keeping purposes, their data might be incomplete. Even when the data is available, it lacks the necessary information and might be inadequate for research purposes due to insufficient information for proper understanding of violence. For instance, the medical record from healthcare organisations contain information about an injury and the medications undertaken, but not the circumstances surrounding an incident. Further, the health records are considered confidential, and they are not readily available for research purposes.

Drawbacks of Survey-based Approaches

Though the survey analysis is considered to be potential source of information, the following limitations can be identified [77, 211]:

- *Time and Cost Intensiveness:* The shortcomings of reliance on survey for information extraction range from the costs and time associated with template design and output processing to potential bias introduced through questions formulation. Additionally, the commonly experienced 'survey fatigue' phenomenon can result in lower response rate or incomplete information provided [111].

- *Sampling Limitation:* The existing survey-based approach depends solely on the sample extracted that carries a risk of inaccurate DV victims population representation. Such shortcoming can lead to a distorted view of an issues prevalent in violence-affected communities, thus directly affecting the help received by the abuse sufferers.
- *Help Seeking Behavior:* As noted by VicHealth [238], the higher safety measures are required in order to handle the abuse-related research. It has been proven that women are reluctant to disclose their suffering, therefore the need for data collection in private space and in the absence of male partners emerged [166].
- *Methodological constraints:* Lack of an automated and standardised approach is both labour and time-intensive. Other limitations include human subjectivity inherent in any manual tasks, reliance on expert knowledge and non-scalability with data growth.
- *Data Compatibility:* Lack of uniformity during the data collection phase is the one of the difficult problems faced by research in DV domain. Given the differences in the way the questionnaires are framed, the data collected vary across organisations. This in turn results in the difficulty, when performing comparisons across communities and countries.

2.6 Real-time Analytics of Social Media Data

The enormous benefit of Social Media is the short time span that the messages can reach wide network of users, thus its major role in real-time analytics. Social Media is also a cost effective and less time-consuming alternative than other

data collection approaches, such as surveys and opinion polls. Due to its ease of use, speed and reach, Social Media has become a platform to set the trends and agendas in topics covering healthcare, politics, technology, stock market analysis, entertainment industry, and so on. As it has become a valuable and convenient source for collective wisdom, many research studies used this power for real-world outcomes prediction.

Social Media produces massive amount of data at an unprecedented scale, thus supports the various types of real-time analytics, e.g. spatial, temporal, or text mining. Spatial analytics can provide the visual representation of trending topics across geographical locations, and temporal analytics is considered useful in obtaining an information about the seasonal trends, or particular topics outbreaks. As an example, Kathy et al. [120] have developed a novel real-time flu and cancer surveillance system that uses spatial, temporal and text mining on Social Media data. The real-time analytics results are reported visually in terms of US disease surveillance maps, distribution of disease types, symptoms, and treatments. In addition, an overall disease activity timelines have been proposed in [120].

Several research studies focused on Twitter for sentiment analysis [12], opinion mining [159, 233], natural disasters [207], epidemic surveillance [40], event detection [183], and so on. Sentiment analysis has been previously studied with respect to different data sources such as blogs and forums, and has now been analysed using Social Media streams [168]. O'Connor et al. [159] and Tumasjan et al. [207] showed that sentiment analysis of online posts correlated with the voters political preferences, and closely aligned with the election results. Not only in the field of politics, but also in economics, have public tweets played a

major role. Bollen et al. [17] [16] analysed that tweets sentiment can be used to predict trends of stocks, and it is directly correlated with them. Bruns et al. [19], Burns et al. [23], and Gaffney et al. [71] have further observed that Social Media is a powerful tool to gather public opinion and induce social change.

Sakaki et al. [207] investigated Social Media data during natural disasters and demonstrated its ability for earthquakes detection and timely warning alerts. They considered each Twitter user as a mobile sensor in Japan, and the probability of an earthquake is computed using time and geo-location information of the user. The volume of posts over time was modelled as the exponential distribution to estimate locations of earthquake using kalman and particle filters. Their research further evidenced that earthquake can be sensed earlier than official broadcasts.

Culotta et al. [49] analysed Social Media to detect influenza epidemic outbreaks, which outperformed the traditional methods in terms of cost and speed. Furthermore, the additional metadata about the users gender, age and location can be used to provide valuable demographic insights, compared to search queries only. The influenza was detected based on multiple regression models. Quincey et al. [57] identified the swine flu from Twitter using pre-defined keywords, and terms co-occurrence. These methods are analysed by searching the tweets with the keywords specified, and investigating any anomalous changes in the flow of messages related to those keywords. The aid of such methods is to curate the actionable information from the unstructured online stream.

Social Media proves to be an effective source of data for research in healthcare-related topics, in particular in various diseases detection and analysis. These include cholera [41], cardiac arrest [18], dental pain [90], alco-

hol use [50], tobacco [44], drug use [175], mood swings [83], or Ebola outbreak [160]. Michael et al. proposed a technique called Ailment Topic Aspect Model [173, 174, 176] to monitor the healthcare of public by the diseases, symptoms and treatments detection in tweets.

This section described the real-time application of Social Media data in various sectors, i.e. healthcare, politics, natural disasters, stock market analysis, sentiment analysis, and so on. The enormous amount of information disseminating through millions of users accounts presents an interesting opportunity to obtain a meaningful insight into the population behavioural patterns, along with the prediction of future trends. Moreover, gathering information on how people converse regarding the particular topic, and distilling the essential information from user-generated content can support numerous real-world applications.

2.7 Chapter Summary

Given the importance of circumstances surrounding violent incidents better understanding, the study proposed Social Media utilisation for the essential information collection. As discussed in previous sub-section (2.5.1), the 'health-related data' (information on diseases, injuries or other health conditions linked to the abusive experience), and 'self-reported data' (information on victimisation, victims/perpetrators characteristics etc.) are considered vital for greater insight into the problem of violence, and more effective preventative strategies development. As a result, this research utilises the abuse reports posted in risk-free online environment, and addresses the limitations of the conventional survey-based approaches. Finally, the work conducted demonstrates that

Social Media platforms serve as a valuable knowledge base for DV prevalence and severity estimation, as the victims increasingly share their personal experiences of abuse in online settings, such as Facebook DVCS groups.

CHAPTER 3

COMPUTATIONAL TECHNIQUES OVERVIEW

The previous chapter discusses about the background of DV and its consequences, role of Social Media in those stigmatized contexts. This chapter discusses about various computational techniques like text mining, Machine Learning and Deep Learning techniques.

Social Media platforms contain wealth of user-generated data and over time has become a virtual treasure trove of information for knowledge discovery with applications in health care, politics, social initiatives, to name a few. Despite the evident benefits of Social Media data exploration, there are numerous challenges associated with processing such data, given specific characteristics of users' posts. This chapter [1] deals with a brief of steps involved in manipulation of such data as well as offers the examples of the Machine Learning and Deep Learning algorithms most commonly used in text analysis. The chapter also provides some exemplary studies that prove the potential of Social Media to play an important role in meaningful results extraction and guidance for decision makers.

3.1 Text Mining for Social Media Application

People generally use their own language to post their views on Twitter or any other Social Media platform. Hence the various types of ambiguities occur, such as lexical, syntactic, and semantic, as the people do not care about spelling and

[1] *The part of this chapter is based on the published work.*
Sudha Subrmani, Sandra Michalska, Hua Wang, Frank Whittaker, and Benjamin Heyward. "Text Mining and Real-Time Analytics of Twitter Data: A Case Study of Australian Hay Fever Prediction." In 7th International Conference on Health Information Science, pp. 134-145. Springer, Cham, 2018.

grammatical usage in the sentence [218]. In recent years, Social Media has become an active research area that has drawn huge attention among the research community for information retrieval and abstract topics discovery. Nonetheless, the following characteristics of Social Media makes it challenging for that purpose:

1. Immense volume, fast arriving rate and short message restriction (especially for Twitter),
2. Large number of spelling and grammatical errors,
3. Use of informal and mixed language,
4. High content of irrelevant data.

Therefore, an extraction of meaningful information from such noisy data became complex problem to solve. Text mining intends to address the above-mentioned issues. Liu et al. [122] defined text mining as an extension of data mining to text data. Text mining differs from traditional content analysis, as text mining techniques focus more on finding hidden patterns and trends in the unstructured data. Whereas data mining techniques are used mainly for knowledge discovery from structured and organised databases [109]. Information retrieval, text analysis, clustering and natural language processing are the multidisciplinary fields in text mining techniques. They facilitate models based on interesting patterns development and assist predictability.

3.2 Pre-processing Steps in Text Mining

During data collection, the unstructured text data contains a lot of challenges as described in previous section. Specific characteristics of Social Media data as described in previous section makes tweets particularly challenging to work with. At the same time, these steps are essential in any subsequent analyses. Precisely, if the data is not cleaned properly, the text analysis techniques at the later stage simply leads to 'garbage in garbage out' phenomena [53]. Even though pre-processing consumes a great amount of time, it improves the final output accuracy [68]. Feature extraction and feature selection are two basic methods of text pre-processing.

The content of collected posts varies from useful and meaningful information to incomprehensible text. The former has people opinion and relevant posts regarding the topic, whereas the latter may contain advertisements and is not worth reading. Hence, high quality information and features are extracted by incorporating some pre-processing techniques. Pre-processing of online stream removes noise that produces negatives effects and degrades performance.

3.2.1 Feature Extraction

The Feature extraction can be further categorized as 3 methods such as 'Morphological analysis, Syntactical analysis and Semantic analysis'. The 3 categories are briefly explained below. The feature extraction is used for many applications likes automatic posts classification [220], opinion analyser from the posts [262] and sentiment classification [216].

Morphological Analysis

This technique mainly deals with 'tokenization, stop-word removal and stemming' [68]. The tokenization is the process of breaking a stream of text in to words or phrases called tokens [148]. Stop word lists contain common English words like articles, prepositions, pronouns, etc. Examples are 'a, an, the, at,' etc. Hasan saif et al. [206] investigated that removing stop words improves the classification accuracy in Social Media data analysis by reducing data sparsity and shrinking the feature space. Stemming is used to identify the root of a word, to remove the suffixes related to a term and to save a memory space. For example, the terms 'relations', 'related', 'relates' can be stemmed to simply 'relate'. Different stemming algorithms are available in the literature, such as brute-force, suffix-stripping, affix-removal, successor variety, and n-grams [68]. Porter stemming [187, 188] is applied to standardise the terms appearance and to reduce data sparseness. In addition to the above 3 methods, non-textual symbols and punctuation marks are removed. Noisy tweets are filtered by eliminating links, non-ascii characters, user mentions, numbers and hashtags

Syntactical Analysis

Syntactic analysis consists of Part-of-Speech tagging (POS tagging) and parsing techniques [256]. It provides knowledge about grammatical formation of the sentence and it is used to interpret logical meaning from the sentence. The POS tagging defines contextually related grammatical sense in a sentence like noun, verb, adjective etc. Various approaches have been developed to implement POS tagging like Hidden Markov Model [256]. Parsing is another technique of syntactical analysis, where the sentence is represented in a tree-like structure and

analysed for which group of words combine.

Semantic Analysis

Semantic analysis is the primary issue for relationship extraction form unstructured text [102]. This refers to wide range of processing techniques that identify and extract entities, facts, attributes, concepts and events to populate metadata fields. This is usually based on two approaches like rule-based matching and Machine Learning approach. First approach is similar to entity extraction and requires the support of one or more vocabularies. Another one is Machine Learning approach and it deals with the statistical analysis of the content and derives relationship from the statistical co-occurrence of terms in the document corpus. WordNet-Affect [222] and SentiWordNet [63] are the popular approaches that are used to extract the useful contents from the textual message. Strapparava et al. [222] proposed the WordNet-Affect approach, a linguistic resource for a lexical representation of affective knowledge (affective computing is advancing as a field that allows a new form of human computer interaction in addition to the use of natural language). Another approach is SentiWordNet, which is proposed by Esuli et al. [63] and it is a publicly available lexical resource for opinion mining.

3.2.2 Feature Selection

Another essential step after feature extraction is feature selection. This improves the scalability and accuracy of the classifier by constructing vector space. The main purpose of this approach is to select the most important subset of features

from the original documents based on the highest score. The highest score is predetermined measure based on the importance of the word [145]. For the text mining, the high dimensionality of the feature space is the major hurdle, as it contains many irrelevant and noisy features. Hence feature selection method is widely used to improve the accuracy and efficiency of the classifier. The selected features provide a good understanding of the data and retain original physical meaning. A substantial amount of research has been applied to evaluate the predictability of features for the application in classification techniques. Among them, Peng et al. [179] studied how to select compact set of superior features at low cost according to a maximal statistical dependency criterion based on mutual information. Another approach is based on conditional mutual information and it is defined as a fast feature selection technique. This approach favours features that maximize their mutual information and ensures the selection of features that are both individually informative and 2-by-2 weakly dependent [66]. Mihalcea et. al. [141] examined several measures to determine semantic similarity between short collections of text. It relies on simple lexical methods like pointwise mutual information and latent semantic analysis.

Another popular approach calculates feature vectors based on two basic methods like Term Frequency (TF) and Inverse Document Frequency (IDF). TF is calculated based how often the term occurs in a collection of documents. The topic information can be determined by TF that is by the number of occurrences of a term associated with the topic. IDF considers the least frequent words in the document that have topic information. Hence, TF-IDF function is the combination of TF and IDF and is mainly used to estimate the frequency and relevancy of a given word in the document at the same time. Ramos et al. examined the results of applying TF-IDF to determine what terms in a corpus of documents

might be more relevant to a query and also performs the efficient categories of relevant words [193]. TF-IDF approach was applied for sentimental analysis of product reviews and improved the computational efficiency [132]. Another study reviewed various techniques used in a classifier to examine the noun-phrase level in Twitter [202]. Language feature sets developed from users social links were applied for feature sets for NP classification in Twitter [39]. Another method of vote-constrained bootstrapping was evaluated in the context of POS tagging. This allowed for improved performance in the application of the POS tagging as applied to users' posts linguistic complexity [59].

3.3 Text Mining Using Classification Techniques

Classification techniques are widely popular for text mining purpose. In general, classification algorithm contains two phases. First, the algorithm learns the model from training dataset for the class attribute from the defined dependent variable. Next, it applies previously trained model on an unseen testing dataset and predicts the class attribute for each instance. Text classification automatically classifies the texts into relevant categories [215]. In the text classification approach using Machine Learning, the classifier learns how to classify the categories of documents based on the features extracted from the set of training data. Thus, the Social Media mining is the combination of data mining and text mining techniques [215]. Various Machine Learning and Deep Learning algorithms are discussed in next sections.

3.4 Machine Learning Algorithms Overview

Some of most popular classification algorithms such as Naive Bayes, Decision Tree, Support Vector Machines, K-Nearest Neighbours, and Logistic Regression are briefly discussed in the following sub-sections.

3.4.1 Naive Bayes

The most generally used classifier for text classification. Basic idea behind this classifier is to estimate the probability to which class the document belongs. Depending on the precise nature of the probability model, the naive bayes classifiers can be trained very efficiently by relatively small amount of training data to estimate the parameters necessary for classification. As the independent variables are assumed, only the variances of the variables for each class need to be determined, not the entire covariance matrix. Due to its apparently oversimplified assumptions, the naive bayes classifiers often work much better in many complex real-world situations than one might expect. It has been reported to perform surprisingly well for many real world classification applications under specific conditions [133], [201]. An advantage of the naive bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification.

3.4.2 Decision Tree

Decision tree classification method uses the rule-based inference to classify documents to their annotated categories [6]. The algorithm built a rule set that describes the profile for each category. Rules are typically defined in the format of "IF condition THEN conclusion", where the condition portion is filled by features of the category and the internal nodes are represented with the categorys name or another rule to be tested. Various categories of splits in the trees are available like Single attribute split, Similarity-based multi-attribute split and Dimensional-based multi-attribute split. Either C4.5 or ID3 can be used for splitting of a node. The tree is designed as hierarchical decomposition of the data space. In order to minimise the overfitting, pruning can be done. In the case of handling a dataset with large number of features for each category, heuristics implementation is recommended to reduce the size of the rule set without compromising the performance of the classification. The advantage of decision rules method for classification tasks is the construction of local dictionary for every individual category during the feature extraction phase [6]. Local dictionaries are able to distinguish the meaning of a particular word for different categories. The disadvantage of the decision tree method is the impossibility to assign a document to a category exclusively due to the rules from different rule sets is applicable to each other. Also, the learning and updating of decision tree methods needs extensive involvement of human experts to construct or update the entire rule sets.

3.4.3 Support Vector Machines

Support vector machines are commonly recognized for their high predictive accuracy. The support vector machines classification method is based on the Structural Risk Minimization principle from computational learning theory [104]. In contrast to other classification methods, the support vector machines need both positive and negative training sets, which are uncommon for other classification methods. These sets are required for the support vector machines to find the decision surface that best separates positive from negative instances of data through linear hyperplane, which maximizes the margin. The document representatives, which are closest to the decision surface are called the Support. The performance of the support vector machines classifier remains unchanged if documents that do not belong to the support vectors are removed from the training set.

3.4.4 K-Nearest Neighbours

The k-nearest neighbours algorithm [88] works on principle that the documents that are close in the space belong to the same class. It is an instant-based learning algorithm used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. The key element of this method is the availability of a similarity measure for identifying neighbours of a particular document and it is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measures. The advantage of this method is its effectiveness, non-parametric property as well

as easy implementation. However the disadvantage is the classification time that is long and it is difficult to find an optimal value of k. It uses all features in distance computation and causes the method computationally intensive, especially when the size of training set grows. Besides, the accuracy of k-nearest neighbours is severely degraded by the presence of noisy or irrelevant features.

3.4.5 Logistic Regression

Logistic regression is a method for binary classification and it is based on the logistic function (also called sigmoid). The two characteristic features of that function makes it particularly convenient for modelling probabilities. These are: 1) it is monotonically increasing 2) its range is between 0-1. As stated before, logistic regression is a probabilistic function, which means that the conditional probability of a data point belonging to a class of interest using the sigmoid function must be fit. The probability of assignment to the opposite class is simply its complement. There are many ways for fitting the best coefficients. In the logistic regression model, the coefficient vector that maximises the joint likelihood of the input data points in the training set having their corresponding label is favoured. As an optimisation technique the gradient descent is most frequently used. What makes logistic regression classifier convenient in text related tasks is the inspection of its coefficients generated from the training set. Given the high level of ambiguity present in all natural language processing tasks (short messages such as tweets in particular), the insight into the classification criteria allows for further algorithm refinement to better fit its purpose. This feature is especially advantageous when the goal is the extraction of relevant data on a particular topic given user defined criteria (e.g. posts using specified key-

words). In that case, both features determination as well as classifier selection and tuning contribute towards overall systems sensitivity

3.5 Deep Learning Algorithms Overview

Deep Learning is a relatively new branch of Machine Learning, whose advantage is the ability to automatically extract intermediate feature representations of raw textual data by building a hierarchical structure [119]. Traditionally, the Machine Learning algorithms efficiency heavily relied on the feature engineering part, and lot of research studies focused on constructing features from raw data, as it requires more domain knowledge and significant human effort. On the contrast, the Deep Learning algorithms requires minimal domain knowledge and human effort, considering the automatic feature extraction of Deep Learning algorithms.

Deep Learning has been applied in various text mining applications, such as sentence modeling [107], text classification [113], and machine translation [139]. Deep Learning also plays a tremendous role in various real-time applications using online Social Media data, which include the detection of cyber-bullying and online harassment [33] [255], disaster response and management [152], massive open online courses forums [247], and personality prediction [125].

Though the application of Deep Learning techniques in other fields are tremendous, the application of Deep Learning on the DV context is unfamiliar and unexplored yet. To the best of our knowledge, no previous work has fully investigated the potential of Deep Learning in this domain. Hence, the popular Deep Learning architectures are discussed in this section, and their real

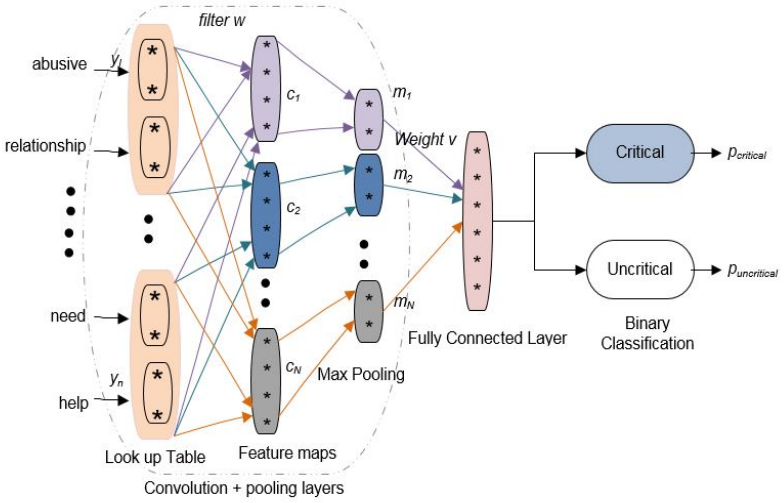


Figure 3.1: Convolutional Neural Network on a Sample Post: "I am suffering with abusive relationship and desperately need help"

time application on various contexts of DV are demonstrated in the upcoming chapters.

3.5.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are the current state-of-the-art model architecture of Deep Learning, which has found applications in text classification problem [113]. CNNs apply a series of filters to the textual features to extract and learn higher-level features, which is later used for classification. Figure 3.1 represents the architecture of CNNs (Source [113]) for an example DV post, "I am suffering with abusive relationship and desperately need help". The architecture is based on the context of critical posts detection, that is explained in the chapter 4 in detail.

Given a input post $\mathtt{P} = \sum_{i=1}^{n} x_i$, where $\mathtt{P} = x_1, x_2...x_n$ are the individual word tokens. We initially transform it into a feature space by mapping each word token $x_i \in \mathtt{P}$ to an index of look-up table L. L can be initialized with random or pretrained word embedding vectors like word2vec or Glove. Look-up table L is represented as $L \in \mathbb{R}^{|V| \times D}$, where each word in vocabulary V is represented by D dimensional vector.The look-up table generates an input vector $y_i \in \mathbb{R}^D$ for each token of word x_i, which passes through sequence of convolution and pooling operations to produce high level abstract features.

Convolution Layer: A convolution operation involves a filter $w \in \mathbb{R}^{|L| \times D}$ to a window of L words to generate a new feature c_i

$$c_i = f(w.y_{i:i+L-1} + b_i) \tag{3.1}$$

where $b_i \in \mathbb{R}$ represents the bias term and $y_{i:i+L-1}$ is the concatenation of L input vectors, and f is a non-linear activation function like tangent. This convolution filter is applied to each possible L window of words in the each facebook post to produce a feature map $c_i = [c_1, c_2..., c_{n+L-1}]$ with $c \in \mathbb{R}^{n+L-1}$. This process is repeated N times with N different filters to obtain N different feature maps.

Pooling Layer: We then apply a max pooling operation over the feature map and takes the maximum value of $m = max\{c\}$ such as

$$m = [\mu_p(c_1), \mu_p(c_2), ..., \mu_p(c_N)] \tag{3.2}$$

Fully Connected Layer: As each convolution and pooling operation is performed independently, the the location of the extracted features become invariant. Hence, to preserve the order information is vital for sentence modeling

of facebook posts. To this end, we included fully connected or dense layer of hidden nodes on top of the pooling layer such as

$$z = f(v.m + b_c) \tag{3.3}$$

where v is weight matrix, b_c is a bias vector, and f is a non linear function. The dense layer deals with sentences of facebook posts, which are of variable length and produce fixed size output vector z, which are given as input for further text classification tasks.

3.5.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) [225] handle a variable-length sequence input by having loops called recurrent hidden state, which captures the information from previous states. At each time stamp, it receives an input, and updates the hidden state. The predictions are accurate, as the hidden state integrates information over previous time stamps. Instead of learning features by traditional neural network models, RNNs has loops called recurrent units, which captures the information from previous states. The decision of recurrent network at time stamp t also relies on the decision made at previous time step $t-1$. Let the input x_t is fed to the network at timestamp t, h_t is the hidden layer captures information from input x_t. The decision of output layer o_t not only depends on current input layer h_t, but also relies on the previous state of h_{t-1}. Hence, the process of carrying memory forward is denoted as,

$$h_t = \phi(Wx_t + Uh_{t-1}) \tag{3.4}$$

Here, ϕ defines the sum of the weighted input and hidden state combined. W is a weight matrix and U is a transition matrix.

When using text data for prediction and to understand the context, preserving information long enough is of paramount significant. Although the RNNs can capture the information from previous states, vanishing gradients become a major obstacle for the higher performance and makes it difficult to learn and tune parameters from previous states [14]. Hence, Gating mechanisms have been developed to lessen the limitations of the conventional RNNs, resulting in two prevailing RNNs types: Long Short-Term Memory networks [84] and Gated Recurrent Units [36]

3.5.3 Long Short-Term Memory networks

The Long Short-Term Memory networks (LSTMs) model [84] is an improvement of the RNNs model by adding a memory component, to add a delay between the input and output. The LSTMs unit adds contextual information to the network and learn long-term dependencies without keeping redundant information. This works remarkably well on Natural Language Processing tasks and sentiment classification in addition to image processing. The LSTMs unit can store the history information by it's memory unit and thus, the input gate, the output gate and the forget gate can be updated by utilizing historical information. The structure of LSTMs unit for our task as follows. First, we compute the values for i_t, the input gate at time t.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{3.5}$$

Second, $\widetilde{C}_t$ the candidate value for the states of the memory cells at time t are computed as:

$$\widetilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3.6}$$

Next, we compute the value for f_t, the activation of the memory cells forget gates at time t:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3.7}$$

We can compute C_t the memory cells new state at time t, when the value of the input gate activation i_t, the forget gate activation f_t and the candidate state value $\widetilde{C}_t$ are given.

$$C_t = i_t * \widetilde{C}_t + f_t * C_{t-1} \tag{3.8}$$

We compute the value of their output gates and the corresponding outputs, with the new state of memory cells.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_t + b_o) \tag{3.9}$$

$$h_t = o_t * \tanh(C_t) \tag{3.10}$$

- Here, x_t is the input to the memory cell layer at time t.
- $W_i, W_f, W_c, W_o, U_i, U_f, U_c, U_o$, and V_o are weight matrices.
- b_i, b_f, b_c, and b_o are bias vectors.

3.5.4 Gated Recurrent Units

The Gated Recurrent Units (GRUs) [36] is basically an LSTMs with 2 gates, whereas LSTMs has 3 gates. GRUs merges the input and forget gates into a single one, named as "update gate". Thus the GRUs is simpler than LSTMs and

has been popular. It modulates the flow of information inside the unit, but it fully exposes the contents to a larger net at each time step without any control.

3.5.5 Bidirectional LSTMs

Another drawback of conventional RNNs is that only historic context can be exploited, where it is helpful to exploit the future context for the typical tagging task. Hence, Bidirectional RNNs (BRNNs) [214] provides an effective solution by accessing both the preceding and succeeding contexts with two separate hidden layers. Initially, it computes the forward hidden sequence followed by backward hidden sequence to generate output. Thus, the Bidirectional LSTMs (BLSTMs) [87] functions by replacing the hidden states in BRNNs with the LSTMs memory units. The BLSTMs layer integrates the long periods of contextual information from both directions.

3.6 Application of Classification Techniques in Social Media Data Analysis

There has been an increasing interest in studies on sentiment analysis, emotional models, opinion mining and topic modelling [169]. It is due to the enormous growth of data available in the Social Media, especially of those that reflect people's opinions, feelings and experiences. Early opinion mining studies focus on document level sentiment analysis and opinion mining from posts published on web pages or blogs [260]. Pang et al. [171] used naive bayes, and support vector machines classifiers to analyse sentiment of movie reviews. They classified

movie reviews as positive or negative and performed the comparison between methods in terms of accuracies achieved. Nowadays, Social Media has received much attention for sentiment analysis and opinion mining as it became a source of massive user-generated content with a wide array of published opinions. The most common approaches such as Machine Learning, Deep Learning and natural language processing techniques address the problem of sentiment analysis and opinion mining on Social Media.

In [250], tweets referring to hollywood movies are classified using naive bayes algorithm. Amolik et al. [4] proposed highly accurate model with respect to latest reviews of upcoming bollywood or hollywood movies using feature vector and classifiers such as support vector machines and naive bayes. Pak et al. [167] performed linguistic analysis of the collected twitter corpus and build a sentiment classifier, which is able to determine positive, negative and neutral sentiments for a document using naive bayes classifier. Similarly, in [78] and [149] authors proposed a method for sentimental analysis of reviews from the Social Media data using Machine Learning techniques like naive bayes and support vector machines. Ye et al. [257] used Deep Learning technique for the sentiment analysis on twitter. In [198], authors proposed a context-based Deep Learning model for the same type of analysis, incorporating contextualized features from relevant tweets into the model in the form of word embedding vectors.

Apart from sentiment analysis, the Machine Learning techniques in Social Media mining have wide range of advantages in biomedical and healthcare fields. Nivedha et al. [157] developed a model to classify the tweets to health and non-health related data using decision tree algorithm. They also concluded

that decision tree classifier performed better than naive bayes for the case tested. In [45], tweets were collected and classified into syndromic categories based on the keywords from the public health ontology. They used naive bayes and support vector machines model to classify the data. Paul et al. [178] classified tweets based on the keywords related to disease, symptoms and their treatment using the ailment topic aspect model. The support vector machines was trained with linear kernel and uni-gram, bi-gram and tri-gram features. In [7], authors used support vector machines algorithm to identify the flu related posts using uni-grams. Ramya et al. [194] applied the classification algorithms C4.5 and support vector machines for the advocacy monitoring from tweets. The extracted women and children health data from Social Media was utilised for advocacy monitoring using classifiers. Similarly, another study proposed automated classification approach for mental health-related posts identification from Social Media, followed by further categorization into the specific disorders such as bipolar, anxiety or depression, based on the underlying theme of the posts using Deep Learning techniques [80].

Not to mention the tremendous advantage of applying Machine Learning and Deep Learning techniques in crisis response situations [3, 30, 74, 98] during natural disaters. Imran et al. [99] used naive bayes algorithm for the automatic categorization of user posts. In the similar study of crisis response, nguyen et al. [151] categorized the Social Media posts using CNNs algorithm.

Bollen et al. analysed how Social Media mood predicts the stock market using Deep Learning methods, as behavioural economics states that emotions can profoundly affect individual behaviour and decision-making [17]. Spam contents in Twitter were found to be distracting and annoying for certain users,

thus mobile application to deliver spam-free Twitter trending topics contents is needed. Hence, Aryo et al. [120] implemented naive bayes and K-nearest neighbours to detect spam in Twitter trending topics.

Thus, this section demonstrated the potential of automatic classification techniques in the various context of Social Media applications.

3.7 Chapter Summary

This chapter provides the high-level overview of various Machine Learning and Deep Learning algo rithms popular in text mining as well as the application of these techniques in the analysis of Social Media data. The numerous examples from previous studies that transform unstructured content into valuable knowledge are given along with the classification techniques implemeted. The brief summary allowed to obtain a good understanding of the available techniques on working with Social Media data, and their application on real-world problem of DV are explained in upcoming chapters.

CHAPTER 4

DOMESTIC VIOLENCE CRISIS IDENTIFICATION FROM SOCIAL MEDIA POSTS BASED ON DEEP LEARNING

The availability of Social Media has allowed DV victims to share their stories and receive support from community, which opens an opportunity for DVCS to actively approach and support DV victims. Hence, there is a need for quick identification of the victims of this condition, so that DVCS groups' moderators can offer necessary support in a timely manner. However, it is time consuming and inefficient to manually browse through a massive number of available posts. This chapter [1] presents Deep Learning based approach for automatic identification of DV victims in critical need. Empirical evidence on a ground truth data set has achieved an accuracy of up to 94%, which outperforms traditional Machine Learning techniques. Analysis of informative features helps to identify important words which might indicate critical posts in the classification process. The experimental results are helpful to researchers and practitioners in developing techniques for identifying and supporting DV victims.

4.1 Introduction

In recent years, social-networking platforms (et. Facebook and Twitter) have exploded as a category of online discourse [74], which has shown their important role in the dissemination of supporting information and providing actionable situational knowledge during crises situations [67]. DVCS has been aware of

[1] *This chapter is based on the following article.*
Sudha Subramani, Hua Wang, Huy Quan Vu, and Gang Li. "Domestic Violence Crisis Identification From Facebook Posts Based on Deep Learning." IEEE Access, vol 6 (2018), pp. 54075-54085.

Table 4.1: Examples of DV Posts and the Corresponding Intent Labels

ID	DV Posts	Context	Label
P_1	"To understand why people stay in abusive relationships, visit the link."	Awareness promotion	Uncritical
P_2	"The code of silence is bulls**t"	Personal opinion	Uncritical
P_3	"Morning greetings. Enjoy the day."	Greetings	Uncritical
P_4	"Rest in Peace, Beautiful Angel."	Expressing empathy	Uncritical
P_5	"I hope that this is okay I desperately need help. Please read my story and consider helping me I am desperate thank you very much."	Shared by victim	Critical
P_6	"My best friend is fighting for her freedom. My dear friend was brutally beaten by her bf."	Shared by acquaintance	Critical
P_7	"A woman who was shot dead by the father of her children on Anzac Day sacrificed her life to save her children."	Shared by media	Critical

the benefit that the Social Media platforms can bring to aid their decision and approach to support victims of DV. An issue with the posts shared on Social Media DVCS is that they are available at large scale, while not all posts are critically important. For examples, the posts P_{1-4} in Table 4.1 are relevant to DV, but they are mainly for promoting awareness, providing advice, or expression of empathy. Such posts can be treated as 'uncritical', as they do not describe a situation where a person is in danger or need immediate support. In contrast, the posts P_{5-7} describe 'critical' situations, where victims may need immediate support from DVCS. The accurate identification of such critical posts are crucially important for DVCS to direct their limited resources to support those in critical need. Manual browsing through a large amount of online posts is time consuming and inefficient to identify critical posts. As such, a tool that can filter the online posts relevant to DV and flag those critical posts is needed.

The use of online posts to support decision making in crisis has been investigated in the literature, such as during natural disasters of floods [30] and earthquakes [151]. By far, no attempt has been made to develop techniques for identifying personal crisis due to family disruption or disturbance in case of DV.

The identification of critical posts relevant to DV is challenging task. The posts are in form of free text, which is unstructured data. How to represent the textual data for effective identification of critical posts is itself a critical task. It is also unknown, which features might provide an important clue to identify critical posts.

To the best of our knowledge, no prior work has either focused on critical post identification from Social Media or evaluated Deep Learning and Machine Learning techniques against different feature extraction methods for DV identification. Hence, this chapter aims to provide support for DVCS by introducing an approach to automatically recognize critical posts on Social Media platforms.

Firstly, a benchmark data set of online posts with labels, 'critical' and 'uncritical', is constructed. Textual features are then extracted from the unstructured textual data for further processing. Deep Learning, a modern and advanced Machine Learning architecture, is then applied to construct prediction models for automatic identification of critical posts. We treat the problem of critical post recognition as a binary text classification task, where a post is classified as 'critical' or 'uncritical' based on the textual content. We evaluate the performance of an introduced approach against various features for textual data and other traditional Machine Learning techniques. Analysis of informative features help to identify important words, which can distinguish between critical and uncritical posts. The experiment results and analysis are beneficial to researchers, who are interested in carrying out further research in DV based on online Social Media data.

The main objectives of this chapter are summarized as follows:

- DV corpora creation ('gold standard') with binary-class annotation ('critical' or 'uncritical');
- Performance comparison of various feature extraction models;
- State-of-the-art Deep Learning models classification accuracies comparison;
- Superior performance of Deep Learning over Machine Learning empirical validation;
- Knowledge discovery from the semantic coherence analysis of DV related words.

The rest of this chapter is organized as follows. Section 4.2 provides the background on other binary text classification problems, relevant textual features, Machine Learning and Deep Learning techniques. Section 4.3 presents an approach for critical post recognition for DV. Section 4.4 provides details on experiments to evaluate our approach with analysis of the results and discussion. Section 4.5 concludes this chapter and envisages future research directions.

4.2 Literature Survey

With the increasing popularity of Social Media, the amount of information and the range of applications now available to decisive moment is enormous. The advantages of Social Media in information dissemination has been used for several applications such as emergency planning, response and recuperation [74] during disasters like earthquake [98], tsunami [3], and flooding [30]. However, the potential advantages of Social Media in identifying and giving moment support for DV victims, who are in critical need, has not been figured it out.

Automatic labelling of online posts as "critical" or "uncritical" is effectively the classification problem of unstructured textual data. This is basically casted as a supervised text classification task that is basically comprised of two main elements, namely *features engineering* and *label prediction*. The Machine Learning algorithms rely on manual feature engineering, which extracts significant vocabulary items from the textual data and represent them in suitable format, for further analysis. As discussed in previous sections 3.2.1, and 3.2.2 some of the widely used features engineering approaches are Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) [209] [208], psycholinguistic features [180], word n-grams [242], topic modeling features [15], syntactic relations [252], semantic features [51], and sentiment lexicon features [114].

The following step, i.e. label prediction, entails the Machine Learning model training with the features extracted on the 'ground truth' annotated data (also known as 'gold standard'. The most optimal model is subsequently applied to predict the class on the unseen dataset. Some of the most popular Machine Learning algorithms [112, 137, 217] for text classification tasks are: Support Vector Machines (SVM), Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Random Forests (RF), k-Nearest Neighbors (KNN). Though, the performance of the aforestated classifiers heavily rely on the quality of the features extracted.

The popular computational Machine Learning approaches as discussed in section 3.4, which have been applied to various tasks for automatic text classification already. Examples of such tasks include cyberbullying prediction and online harassment [199] [52] [51], emergency situational awareness and crisis response [253], emotion detection and sentiment analysis [147], and opinion min-

ing [170].

The literature work in the areas of aggression, hate speech, disaster crisis response, and mental health are discussed as follows. Simple textual features such as BoW, word and character n-grams were considered to be influential predictors in hate speech detection or abusive language prediction tasks on Social Media [22, 58, 138, 158, 243–245]. For automatic prediction of language posted on Social Media as 'abusive or non-abusive', nobata et al. [158] developed a Machine Learning based method and experimented with various features such as n-grams, linguistic, semantic and syntactic features. The results shown that, character n-grams performed well in the noisy datasets of Social Media. In the another similar study of hate speech binary class prediction as 'hate or no hate', vignal et al. [58] experimented with the SVM model leveraging various features such as syntactic, sentiment, and lexicon features and achieved higher accuracy. In the context of automatic classification of disaster related posts and accurate information extraction, the studies [100,236] used Machine Learning algorithms with comprehensive set of features and achieved higher performance.

In the domain of crisis response during disaster, some works [8, 29, 95] in disaster posts classification used the standard BoW feature model based on unigrams/bigrams, and achieved higher performance. According to Verma et al., [236] maximum entropy algorithm performed well in identifying the situation awareness posts of disaster across the four events of disaster, compared to NB classifier. On the similar study by Imran et al. [100], RF classifier achieved good results.

Linguistic Inquiry and Word Count (LIWC) [180] has been commonly used to capture language characteristics and has shown to be influential features

for predicting depression-related disorders and mental health conditions [192] [221]. For predicting depression, linguistic styles such as an expression of sadness or the use of swear words have been used as the cues [203]. Thus, for predicting mental health conditions, the textual and psycholinguistic features were analyzed [153–155]. Balani et al. [11] used standard n-gram features, submission length and author attributes to classify a mental health disclosure as higher or lower levels of self-disclosures. Popular bayesian probabilistic modelling tools, such as Latent Dirichlet allocation (LDA) were used to extract the topics [94]. LDA and its variants have been used previously to discover several mental ailments discussed in the millions of tweets [177].

However, the performance of the aforestated classifiers heavily rely on the quality of the features extracted. The popular features used for training the Machine Learning models, such as BoW and TF-IDF, prove ineffective due to inherent over-sparsity and non-semantic representation [208]. As an example, the terms 'physical violence', 'physical abuse' and 'physical assault' would be treated as separate features, although they share similar meaning in the context of DV. Semantic relationships between the terms are lost if the traditional manual features engineering is considered. To account for such shortcoming, the state-of-the-art Deep Learning approach is used in order to capture the words dependencies such as synonyms, misspellings and abbreviations, commonly found on Social Media, and resulting in the substantial classification performance improvement.

Deep Learning is a relatively new branch of Machine Learning, whose advantage is the ability to automatically extract intermediate feature representations of raw textual data by building a hierarchical structure [119]. Deep

Learning has been applied in various Natural Language Processing applications, such as sentence modelling [107], text classification [113], and topic categorization [105]. Deep Learning also plays a tremendous role in various real-time applications using online Social Media data, which include the detection of cyber-bullying and online harassment [10,73], disaster response and management [30,150], and massive open online courses forums [247].

As discussed in the previous section, there are two primary Deep Learning architectures, CNNs [113] and RNNs [225]. Both these models take input as the embedding of words in the text sequence, and generates the real-valued and continuous feature vector for the words. CNNs has been applied in sentence-level sentiment classification and and question classification [107, 113] which show advanced performance over traditional Machine Learning techniques (SVM, MaxEntropies). Similarly, RNNs are implemented to model the text sequence and achieved improved performance for multi-task learning [124]. The improved version of RNNs such as LSTMs [84], GRUs [36], and BLSTMs [85] are widely used in Natural Language Processing applications due to their long range dependencies and storing historical information over time.

CNNs were used to classify tweets into different categories such as hateful (racism, and sexism) vs non-hateful speech and outperformed LR classifiers with high precision [73]. For the similar task conducted in [10], LSTMs proved significantly superior to the CNNs and traditional methods such as LR and SVM. In the context of crisis management and response during natural disasters (earthquake [150], and flood [30]), CNNs were adopted to classify the Social Media posts as either informative or non-informative, and resulted in improved performance over the traditional classifiers such as SVM, LR and RF. In the other

study regarding online posts classification in the emergency situations, RNNs outperformed the CNNs and SVM [186]. For various Natural Language Processing applications such as sentiment analysis and question-answering [254], GRUs and LSTMs proved superior over the CNNs. In the example of Russian tweets sentiment classification, GRUs achieved higher accuracy than LSTMs and CNNs [232].

Furthermore, the promising results of Deep Learning techniques are increasingly observed in numerous real-time Social Media applications. These include abusive language towards racism and sexism detection [10], aand other aggressive posts prediction [200]. Nonetheless, all of the evaluated Deep Learning models demonstrate superior text classification performance, yielding comparable results. Still, the selection of the most optimal model is highly dependent on the application task as well as the hyper-parameters setting.

However, no attempts has been made to investigate the potential of deep learning in applications of DV context. Inspired by the above-mentioned works, and to identify the best performing features in DV crisis prediction, the various sets of features extracted from the users' posts were compared. Further, the traditional Machine Learning algorithms with BoW, TF-IDF and LIWC features were comapred with the Deep Learning models that use pre-trained word embeddings, to understand which model combination provide improved classification results. This chapter aims to address the challenges in DV crisis identification by identifying the best performing model for the classification problem of 'critical' and 'uncritical' online posts.

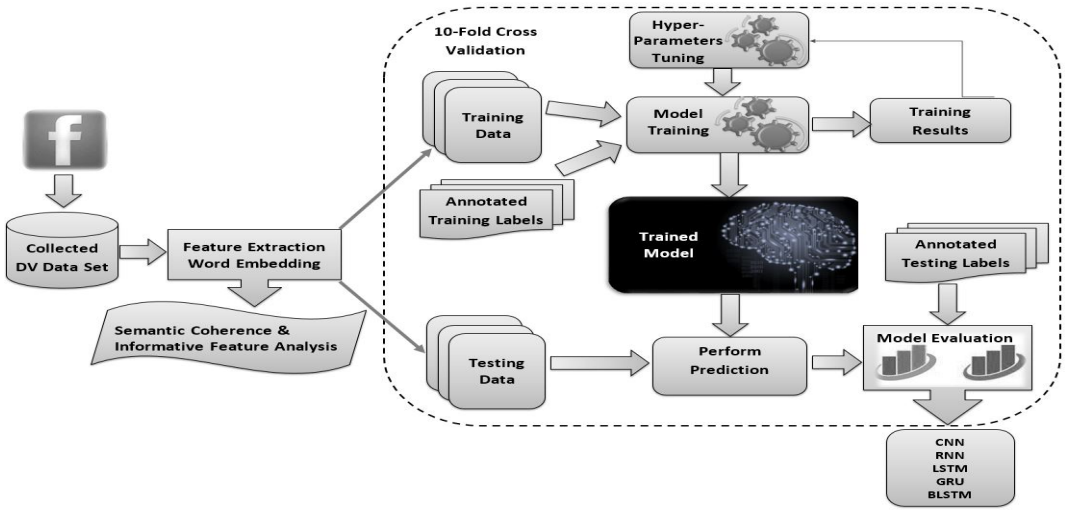


Figure 4.1: Architecture of our Proposed Methodology for Intent Classification of Critical and Uncritical Posts using Deep Learning Model

4.3 Methodology

This section presents our approach (Figure 4.1) to critical post identification, which consists of five stages: 1) Data Extraction: 2) Data Labeling; 3) Feature Extraction; 4) Model Construction; 5) Performance Evaluation. Their details are described in the following subsections.

4.3.1 Data Extraction

Our approach is designed for identifying online critical posts, thus the main source for data extraction is Social Media platforms. We use Facebook as an example to describe the data extraction process, since Facebook is one popular Social Media with around 2 billion users worldwide and ranked first among the top 15 social networking sites [191]. According to [231], emotional support in online context complements the emotional support received in off-line contexts.

Facebook users are benefited by receiving support-based needs (emotional and informational support), due to the ease of sharing with the wide range of people through DVCS. Thus, we collected the posts from pages, that discuss the range of DV issues, through Facebook Graph API [2] with the search term of 'Domestic Violence and Domestic Abuse'. Considering the ethical concern, we collected from the open pages rather than closed and secret pages. The benefit of Facebook Graph API is that researcher can develop applications to detect new posts about DV in real-time, which can support DVCS in quickly identifying DV victims. Please be noted that only publicly available data on Facebook are extracted, which comply with the privacy policy of Facebook. The identity of individuals included in the collected data set are not disclosed in this thesis.

4.3.2 Data Labeling

Our next stage is to label the collected posts as 'critical' or 'uncritical' to construct a benchmark data [3] for evaluating the proposed approach. The posts were manually examined by human scorers independently. If the content of a post was found to describe a critical situation or a situation where a victim indicates the need for help (eg. posts P_{5-7} in Table 4.1), the post is labeled as 'critical'. Otherwise, it is labeled as 'uncritical'. Posts that contain only hyperlinks are treated as irrelevant and discarded from further processing. There is some borderline posts, which may be perceived differently by different human scorers. For example, the post *"I've been there as DV victim. I conquered, I lost my child, I rose up, I walked away, I won that battle scar."* can be treated as 'critical', because it implies that victim needs emotional support, as the child was lost.

[2] https://developers.facebook.com/docs/graph-api
[3] https://github.com/sudhasmani/DV_Dataset

Other scorer may perceive this post as 'uncritical', because he/she reasons that the victim has already battled the situation and that post conveys an aspiration message to stay strong rather than describing a critical situation. The limited context from these posts makes it difficult to interpret fully, and may causes discrepancy in human annotation. As such, only posts that were scored with the same label by all scorers are kept in a benchmark data set for evaluating the Deep Learning algorithms in the later stages.

4.3.3 Feature Extraction

The next stage is to extract features to mathematically describe the characteristics of the data set based on Word2Vec model. The vectors are learnt in such a way that words have similar meanings will have nearby representations in the vector space. Thus, this model overcomes the limitations of the traditional text feature representation techniques such as non-semantic representation and data sparsity. Word2Vec model is the more expressive text representation form, where the relationship between words are highly preserved.

More specifically, Word2Vec takes a textual data as input and each word in the vocabulary is projected as a low dimensional, real-valued and continuous vector, also known as word embedding [143] in the high dimensional space. Suppose an input post is denoted as $\mathtt{P} = \{x_1, x_2...x_n\}$, where x_i is an individual word token in the post $\mathtt{P}$. We initially transform it into a feature space by mapping each word token $x_i \in \mathtt{P}$ to an index of embedding matrix L. Thus, the word embedding matrix is represented as $L_x \in \mathbb{R}^{D \times |V|}$, where D is the dimensional word vector and $|V|$ is vocabulary size. L can be randomly initialized

from a uniform distribution or pre-trained from text corpus with embedding learning algorithms [142] [181]. In simple terms, the mathematical equation represents that the embedding matrix L to be built for each index of the unique tokens in the vocabulary set. We used the latter strategy to make better use of semantic and grammatical associations of words, that is already pre-trained on large external corpus such as Google's Word2Vec [142] and Twitter's crawl of GloVe [181] for our intent classification task.

4.3.4 Model Construction

This stage constructs the prediction model for critical post recognition. We adopt five Deep Learning models for our task, namely:

- CNNs: We adopt the CNNs architecture as described in [105] and used for our approach. Its first layer is called the embedding layer, which extracts the most informative n-grams features and stores the word embeddings for each word. Convolutional layer of CNNs has varying number of computation units, with each unit represents an n-gram (also known as region size) from the input text. Suppose the vocabulary includes V = *'hope','I','was','abused','love'*, there is a post P = *"I was abused"*. In case, the region size is set to 1. The post P is represented as word embedding features [0 1 0 0 0 | 0 0 1 0 0 | 0 0 0 1 0], which is equivalent to Unigram approach. If the region size is set to 2, the post P is represented as [0 1 1 0 0 | 0 0 1 1 0] for the pairs of words *'I was'* and *'was abused'*. This is equivalent to the Bigram approach. Given the variable sizes of the convolutional layer outputs, the pooling layer transforms the previous convolutional repre-

sentation into a higher level of abstract view and produce fixed size output. Finally, the dense layer takes the combinations of produced feature vectors as input and makes prediction for corresponding post. When the consecutive words are given as input, CNNs can learn the embedding of text regions internally, which captures the semantic coherence information in the text.

- RNNs: The RNNs architectures described in [225] is adopted into our approach. RNNs handle a variable-length sequence input by having loops called recurrent hidden state, which captures the information from previous states. At each time stamp, it receives an input and updates the hidden state. The advantage of RNNs is that the hidden state integrates information over previous time stamps.
- LSTMs, GRUs and BLSTMs: LSTMs [84], GRUs [36] and BLSTMs [85] are improved version of RNNs. The core idea behind LSTMs are memory units, which maintain historical information over time, and the nonlinear gating units regulating the information flow. GRUs are basically, an LSTMs with two gates, whereas LSTMs has three gates. GRUs merges the input and forget gates into one unit, named as 'update gate'. BLSTMs consists of two LSTMs, that integrates the long periods of contextual information from both forward and backward directions at a specific time frame. This enables the hidden state to store both the historical and future information. Thus LSTMs, GRUs and BLSTMs are the state-of-the-art semantic composition models for the text classification task and learn long-term dependencies between the words in a sequence, without keeping redundant information.

The models are trained on feature sets extracted from the constructed data

set, so that they can be used to predict the posts as critical or uncritical. In order to examine and compare the prediction performance of the models, we adopt several evaluation measures as presented in the next subsection.

4.3.5 Performance Evaluation

The last stage is to evaluate the performance of the proposed approach to identifying critical posts in relevant to DV. We adopt Precision, Recall, F-Measure, and Accuracy as evaluation metrics of our classifier. They are defined as follows [190]:

$$Precision\,(P) = \frac{TP}{TP + FP} \tag{4.1}$$

$$Recall\,(R) = \frac{TP}{TP + FN} \tag{4.2}$$

$$F - Measure = 2\frac{PR}{P + R} \tag{4.3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

where TP and TN stand for True Positive and True Negative, which measures the numbers of post classified correctly as 'critical' and 'uncritical' respectively. FP and FN stand for False Positive and False Negative, which measures the number of posts classified incorrectly as 'critical' and 'uncritical'.

These metrics have been used widely in various works to evaluate classifier performance [10] [150] [30], which is suitable for our problem of critical post identification. Since, only one data set is constructed for critical and uncritical posts identification, we adopt k-fold cross validation approach for the evaluation. The collected data set is randomly divided into k partitions, where one partition is reserved as test set while the others are combined into a training set. The procedure is repeated k times for different test sets, whose results are averaged to indicate an overall performance.

4.4 Experiment and Analysis

4.4.1 Experiment Design

We start with data collection, where online posts are extracted from Facebook user pages with the keywords 'domestic violence' and "domestic abuse" using its Graph API. A large number of posts and comments were returned. The next step was to label the posts as 'critical' and 'uncritical' to construct a benchmark data set for evaluating the performance of the proposed approach. Since, the labelling process was done manually which is time consuming; we randomly selected a subset of the returned Facebook posts for benchmark data construction. We excluded the posts containing only hyper-links or having less than three words, as they are unlikely to describe a DV situation. The remaining posts are labeled by three research students, under the supervision of a consultant psychiatrist dealing with DV and gender related issues in psychiatric illness, anxiety and depressive illness. The involvement of the domain expert is necessary to

ensure the quality of the labeled data set. We used Kappa coefficient [136] to validate the inter-rater reliability of the human scorers. The achieved degree of agreement was reasonably high at 0.85. Only posts that have consistent labels by all scorers were included in the final data set. We arrive with 750 posts with label 'critical' and 1310 posts with label 'uncritical'. This is a data set with considerable size, considering no previous work on identifying DV victims in critical needs from Social Media data was carried out.

Several experiments were performed to evaluate the performance of the introduced approach using Deep Learning, namely:

(a) *Psycholinguistic Features Analysis:* The textual features from the users posts were also extracted using psycholinguistic knowledge based on LIWC [180]. These features are capable to differentiate the semantic-syntactic patterns and informational context of two different classes of 'critical' and 'uncritical'. The extracted features were experimented with the Machine Learning classifiers and the performance of the classifiers were compared.

(b) *Accuracy Evaluation:* We evaluate the performance of five Deep Learning models, CNNs, RNNs, LSTMs, GRUs and BLSTMs on the constructed benchmark data set. Additional experiments using traditional Machine Learning techniques, with the BoW and TF-IDF features were also carried out for comparison purpose. We compared the performance of classifiers using various evaluation metrics such as Precision, Recall, F-Measure and Accuracy.

(c) *Hyper-parameters Evaluation:* The performance of Deep Learning models can be influenced by their associated hyper-parameters, such as pre-trained word embeddings, selection of optimizer, dropout rate, number

of recurrent units, and number of LSTM memory units or convolution filters. Thus, we carried out experiments with various hyper-parameters to examine their influence to the classification performance. Since, training and tuning a neural network can be time consuming [96] [97], some of the most important parameters, based on the study by reimer et al. [196] were selected for evaluation.

(d) *Semantic Coherence Analysis:* We first examine some important words that may help to distinguish posts belonging to different classes. Then, we examine the semantic composition of the textual features generated by word embedding. The analysis demonstrates the ability to capture semantic meanings between words results in better prediction performance for the Deep Learning models.

In the above experiment, the features for Deep Learning model was extracted using pre-trained Word2Vec models. We used the pre-trained models on two different data sets, Google News [142] and general Twitter posts [181], to examine the robustness of the algorithms. Word2Vec features trained on Google news includes 300 dimensional vectors for a vocabulary of 3 million words and phrases that trained on roughly 100 billion words. Word2Vec features trained on Twitter posts includes 300 dimensional vectors for a vocabulary set of 2.2 million words and phrases that trained on roughly 840 billion words. Thus, for both feature sets, each word is represented by a vector of word embedding containing $D = 300$ dimensions. The first layer of the models is the embedding layer that computes the index mapping for all the words in the vocabulary, and then convert into dense vectors of fixed size by parsing the pre-trained embedding. The next layers contain 128 memory cells, which is popularity used in various applications [196]. The models were trained up to 50 epochs and implemented

using Keras [37].

For the traditional class models, we used TF-IDF and BoW features, because they have been widely used in various text classification applications. We considered 3 different cases of preprocessing for these features, which include (a) stop-words removal only; (b) stemming only; (c) both stop-words removal and stemming, because, the traditional Machine Learning techniques may produce different results with different settings. Average numbers of words in a post before pre-processing, after stop-words removal and after stemming are 155, 71 and 151 respectively. For Deep Learning models, the pre-processing is not carried out, because Deep Learning models process the sequence of words in the order they appear. Stop-words might hold valuable information that could be leveraged. Words are preserved in their original form without stemming, as they can represent different context (e.g. the words 'abusive', 'abuser', 'abuse' are context dependent). For the Deep Learning models, Nadam optimizer is used. Batch size was set to 32 posts, as the dataset size was moderate. Relu activation function and recurrent units set to 128 was used.

4.4.2 Psycholinguistic Features Analysis

This section first examines the features extracted based on the proportions of word usage in psycholinguistic categories as defined in the LIWC 2015 package [180] and then analyses the classification performance. The LIWC analyses text on a word-by-word basis and calculates the percentages of words that match particular word categories. LIWC package is a psycholinguistic lexicon created by psychologists with focus on identifying the various emotional, cognitive, and

linguistic components present in individuals' verbal or written communication. For each input of a post, it returns more than 70 output variables with higher level hierarchy of psycholinguistic features such as

- linguistic dimensions, other grammar, informal language
- temporal, affective, social processes
- cognitive, biological, perceptual processes
- personal concerns, drives, relativity.

These higher level categories are further specialized in sub-categories such as in

- biological processes - body, sexual, health and ingestion.
- affective processes - positive emotion, negative emotion and negative emotion further sub-classified as anger, anxiety, and sadness.
- drives - affiliation, acheivement, power, regard, and risk.

For evaluating the prediction accuracy of psycholinguistic features, each individual post is converted to a vector of 70 output numerical variables, as mentioned above. Each output variable represents the frequency distribution of the appearance of those categories appeared in the specific post. Each word in the post could fit some categories and not fit into some categories. Hence, there would be the huge difference between the posts, to which category it belong. For instance, the following post "*Please view, share and is possible donate. We appreciate your support!*" has higher value of *'positive emotion (36.36%), focus present (36.36%), you (9.09%), social (27.27%)'* and has *(0%)* for the categories such as *'negative emotion, shehe, bio, body'*. The above post falls into 'uncritical' category,

as it creates a good social cause of awarness promotion, and also has higher percentage of positive expression and present focus in it. In contrast, the post "*He is just an evil, greedy, arrogant little man*" has higher percentage of *'negative emotion (40%), anger (10%) , male (20%), shehe (10%)* and *(0%)* of *'posemo, you, death'*. This post falls into 'critical' class, as victim explained about her abusive partner and the post carried a negative emotion in it.

Most Informative Features and Performance Analysis

The most informative 15 features were selected based on the guidance from the domain expert, as shown in Table 4.2 to perform the binary classification task. The Table 4.3 further shows the differences in the mean value of the posts of two different classes. LIWC sub-categories such as *'negative emotion, anger, shehe, focuspast'* are features with higher mean value and good prediction level for 'critical' class. *'Positive emotion, focus present, you, focus future'* sub-categories, as expected, are good predictors for 'uncritical' class. Another important prediction is that *'health, sexual issues'*, and personal concern such as *'death'* are as good predictors for 'critical' class. The results infer that, because of the abusive cycle, most of the victims suffer from severe health issues. Further analysis show that posts related to 'critical' class are often self-reflective, with more words related to personal pronouns i.e, usage of more pronouns such as *'I and shehe'*, when describing their life experience about violence, whereas in the 'uncritical' class, 2^{nd} person usage *'you'* is higher, when giving advice or sharing opinion to other people. It is important to compare the time orientations, the posts of 'critical' category are more focused on *'past'* and contains negative emotions with expression of *'anxiety, angry and sad'*. On the other hand, the 'uncritical' class contains

Table 4.2: LIWC Features and the Sample Words used in the Dataset

Category	Dimension	Example words
Linguistic Dimensions	personal pronouns(I,you,shehe)	I, you, he, she, his, him, her, herself
Time orientations	focuspast	broke, ran, accepted
	focuspresent	supports, trust, likes
	focusfuture	plan, wish, hopeful
Biological Processes	body	muscles, injury, fat
	sexual	rape, lust, abortion, pregnant
	health	sick, weak, painful, bleed
Psychological Processes	posemo	hope, share, support, like
	negemo	threat, lose, hate
	anxiety	threat, misery, worry
	anger	sucks, hate, yell
	sad	miss, lose, suffer, overwhelm
Personal Concern	death	die, murder, kill, suicide, bury

Table 4.3: Mean Scores of Psycholinguistic Features for 2 Classes

Features	Critical	Uncritical	Features	Critical	Uncritical
I	3.15	2.02	Health	1.01	0.61
You	0.59	4.01	Death	1.19	0.02
Shehe	10.98	0.39	Posemo	2.57	10.73
Focuspast	7.09	0.10	Negemo	4.78	1.55
Focuspresent	7.11	14.34	Anxiety	0.65	0.09
Focusfuture	0.96	1.55	Anger	2.28	0.57
Body	0.84	0.42	Sad	0.58	0.14
Sexual	0.34	0.09			

more of positive emotion, as sharing of good thoughts and opinions and more time orientated towards *'present and future'*.

Parallel coordinates plot as shown in Figure 4.2, with all the selected features that separates the class value best. For example, *'posemo, focuspresent, you'* are the features best classifying the class to be in 'uncritical', which is plotted in blue color. The orange color plot explains the class to be in 'critical', with the selective features such as *'shehe, focuspast and death'*. This can be interpreted as, when the victim or survivor posts about abusive experience, they use more past tense and health concern. 'Shehe' notion also widely used to represent

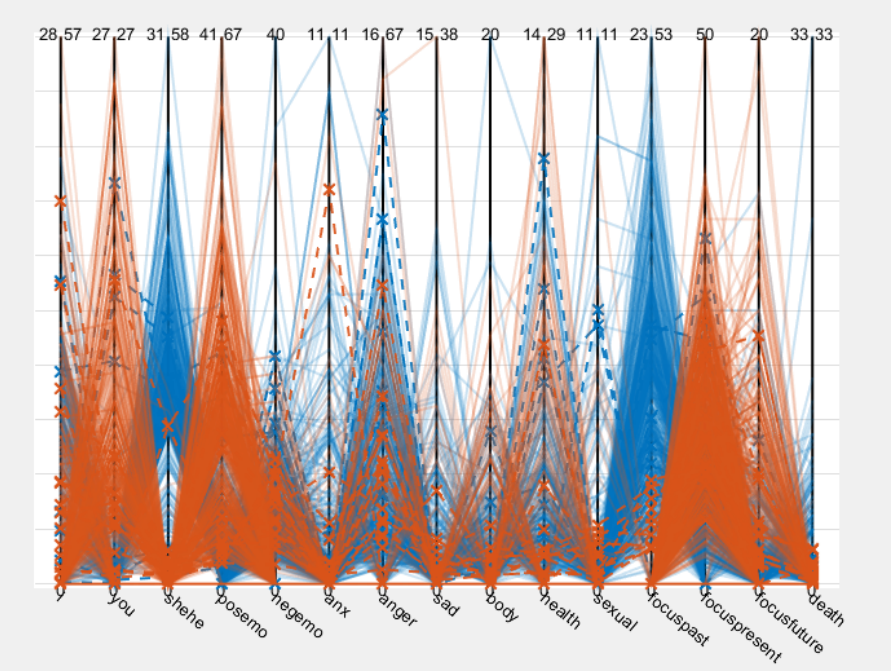


Figure 4.2: Parallel Coordinates Plot for Critical and Uncritical Posts

the abuser. Whereas, in the case of 'uncritical' class, the linguistic style contain present tense and future tense, as it is more focused on future life and well-being.

The Machine Learning algorithms were constructed to predict the classes of the posts based on the features extracted. The chosen most informative features of LIWC have higher accuracy in prediction of two different classes 'critical and uncritical'. Among all the machine classifiers compared, SVM classifier achieved higher accuracy of upto 97% for the 15 features identified manually. Table 4.4 show the various evaluation metrics of SVM classifier with the combination of various selected features. Despite the higher accuracy, the classification performance heavily relied on the feature engineering part, as it requires more domain knowledge and significant human effort. To overcome the limitations, the Deep Learning approach with word embeddings is proposed in this work. Deep Learning approach eliminates manual feature engineering effort,

Table 4.4: Performance Evaluation of LIWC Features

SVM Classifier features set	Precision	Recall	F-Measure	Accuracy
All features (70 features)	73	71	72	76
Linguistic Dimensions (3 features)	96	91	93	94
Time orientations (3 features)	98	86	91	92
Biological Processes + personal concern (4 features)	89	42	57	68
Psychological Processes (5 features)	83	86	84	84
Selected LIWC features (15 features)	97	96	96	97

facilitating more automated and systematic approach are discussed in the next sub-section.

4.4.3 Accuracy Evaluation

The Machine Learning algorithms with traditional textual features were applied to the constructed data set. Since, a single data set were constructed for evaluation, we partitioned the data into training and test sets following 10-fold cross validation approach to measure the performance of the algorithms. We first evaluated the traditional classifiers with different word settings to identify the best setting for comparing with the Deep Learning classifiers. The results as shown in Table 4.5 indicate that the traditional classifiers achieved the best performance with stemming only setting. In the context of DV crisis identification, some stop-words could be helpful to distinguish critical and non-critical posts. We used the stemming only setting for traditional classifiers to compare with other Deep Learning techniques. Evaluation metrics, precisions, recall, and accuracy were computed, as shown in Table 4.6.

In general, Deep Learning models, except for RNNs, achieved better performance than traditional Machine Learning techniques, as indicated by higher

Table 4.5: Accuracy of Machine Learning Classifiers with Different Preprocessing Steps

Classifiers	Stop-words removal only	Stemming only	Both stop-words removal and stemming
NB+TF.IDF	89.65	91.66	89.55
SVM+TF.IDF	91.42	92.20	91.61
RF+TF.IDF	86.85	89.24	87.44
LR+TF.IDF	88.08	90.74	89.26
DT+TF.IDF	86.27	88.58	86.07
NB+BoW	86.12	87.40	84.36
SVM+BoW	89.60	90.82	89.75
RF+BoW	80.19	84.97	82.44
LR+BoW	89.26	90.72	89.94
DT+BoW	86.61	89.83	85.53

evaluation metrics, that showed lower performance. RNNs achieved lower performance among all Deep Learning Models and probably due to the problem of vanishing gradients. Given a long sequence, information of initial sequence fades away as the new sequences are fed into the networks of RNNs. Nevertheless, such limitation of RNNs seems to be overcome by its later versions LSTMs, GRUs and BLSTMs. These models can capture long term dependencies efficiently, which is suitable for dealing with sequential textual data. The Deep Learning models appear to achieve better performance with GloVe than with Word2Vec. With Word2Vec embedding, LSTMs achieved best performance of 93.08%. With the GloVe embedding, LSTMs, GRUs and BLSTMs achieved relatively similar accuracy of more than 94%, which is better than all other algorithms.

Deep Learning approach has been adopted in order to account for the limitations of the conventional Machine Learning techniques in accurate identification of non-standard expressions from Social Media, in the context of DV. To further gain the deeper insights into how the Deep Learning model benefits the classification performance with respect to critical posts identification, some of the posts from the test dataset were sampled in Table 4.7. The examples given are

Table 4.6: Evaluation Metrics of Classification Models

Model	Precision	Recall	F-Measure	Accuracy
CNNs+Word2Vec	92.40	92.40	92.40	92.30
RNNs+Word2Vec	76.90	77.20	77.05	77.09
LSTMs+Word2Vec	93.30	93.30	93.30	93.08
GRUs+Word2Vec	92.70	92.70	92.70	92.64
BLSTMs+Word2Vec	92.80	92.80	92.80	92.54
CNNs+GloVe	93.90	93.90	93.90	93.82
RNNs+GloVe	86.00	85.70	85.85	85.72
LSTMs+GloVe	94.10	94.10	94.10	94.02
GRUs+GloVe	94.50	94.50	94.50	94.26
BLSTMs+GloVe	94.40	94.40	94.40	94.16
NB+TF-IDF	91.80	89.60	90.70	91.66
SVM+TF-IDF	92.90	91.10	92.00	92.20
RF+TF-IDF	90.60	87.20	88.90	89.24
LR+TF-IDF	90.80	89.80	90.30	90.74
DT+TF-IDF	89.15	87.90	88.53	88.58
NB+BoW	87.90	86.50	87.20	87.40
SVM+BoW	91.80	89.80	90.80	90.82
RF+BoW	85.00	83.80	84.40	84.97
LR+BoW	90.80	89.80	90.30	90.72
DT+BoW	89.80	88.80	89.30	89.83

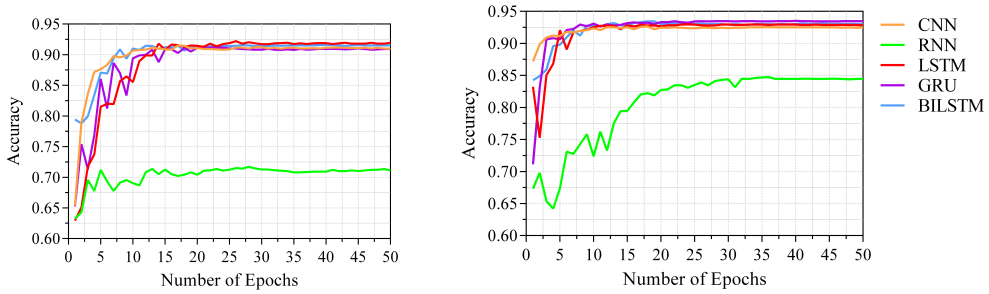
based on GRUs model, as the maximum classification accuracy was achieved by that model with pre-trained GloVe embeddings of 300 dimensions (94.26%). The example posts with accurate predicted classes and the corresponding prediction probaility demonstrates the effectiveness of Deep Learning approach. The posts with prediction probability of less than 0.5 were classified as uncritical, whereas the posts with the prediction probability of more than or equal to 0.5 were classified as critical class. Thus, the posts (Table 4.7) were classified as uncritical with minimum probability of nearly 0.01%, whereas the critical posts were predicted with the higher probability of nearly 0.97% on average.

4.4.4 Hyper-parameters Evaluation

We first evaluated the performance of the Deep Learning models with respect to training epochs. Ideally, the more training epochs would result in well-trained

Table 4.7: Examples of DV posts and the Corresponding Predicted Labels

ID	Posts	Predicted_Class	Probability
P_1	*"Love yourself first always know you deserve to be loved and respected every single day "*	Uncritical	0.030
P_2	*"October is domestic violence awareness month wear a purple ribbon to support "*	Uncritical	0.009
P_3	*"I have overcome dometic violence the loss of a child and a stroke and also much much more. Please read and get some inspiration from my book. Thank you "*	Uncritical	0.063
P_4	*"Hey everyone. I have started a new support group for domestic violence survivors. I am also writing a book about domestic violence. I would love to add a few of your stories to my book feel free to join"*	Uncritical	0.004
P_5	*"Empowering domestic violence survivors to accelerate their own healing process "*	Uncritical	0.014
P_6	*"People who say they are there for you but they actually arent. Its ok to feel disappointed. Its the truth the truth hurts perhaps they are doing you a favor x"*	Uncritical	0.010
P_7	*"Domestic violence is not fun. Survivors of domestic violence need a positive support system." "*	Uncritical	0.012
P_8	*"I personally experienced abuse as a child and carried some of that scaring into my adulthood causing me to act out in similar ways towards my children and husband. There is a saying which i believe to be true. Hurt people we don't intend to but it is part of the by product of abuse. When we don't take the time to get healed sometimes as was the case with me. we know something has to change but we don't know how to change it. "*	Critical	0.968
P_9	*"People say how can you be abused there is no scars they dont see the emotional part of being abused what is does to you. I left my husband four years ago i haven't written about it talked to much about it you dont realize you are being abused well at least. I didn't know there was emotional abuse verbal abuse anything like that til someone close to me pointed it out i was being controlled i guess fear of going out somewhere..."*	Critical	0.982
P_10	*"I knew when i married my husband that there will be problems and dv yet i had no choice. I had to build up strength through the marriage to fight the sexual abuse from my stepfather, the violence continues now coming from the system trying to stay positive through all of this is a hard and very bitter sweet story through my studies. I am hoping to put a mirror in front of all private and system offenders..."*	Critical	0.989



(a) Word2Vec (b) GloVe

Figure 4.3: Accuracy of Deep Learning Models at Different Epoches

and stable models. However, Deep Learning models often take a long time to run. Setting high number of training epochs would result in significant and unnecessary costs. Figure 4.3 shows the accuracy of the Deep Learning models on the two feature sets (Word2Vec and GloVe) with respect to various training epochs. The models appear to converge faster on GloVe features set than on Word2Vec feature set. With Word2Vec embedding (Figure 4.3a), the accuracy of Deep Learning models fluctuated at the beginning and then become stable at their performance after 30 epochs on average. With GloVe embedding (Figure 4.3b), most models become stable after 20 to 23 epochs, except RNNs. Thus, Deep Learning models attained the optimal accuracy and consistency in learning rate, in minimal training epochs with respect to GLoVe embedding.

Next, we evaluated the performance of Deep Learning models with different hype-parameters settings, including optimizer, batch size, number of recurrent units, and activation function. We focused on evaluating GRUs and LSTMs, as they achieved highest performance as shown in the previous sections. The accuracies with 10-fold cross validation are shown in Table 4.8.

Among the optimizers, SGD is quite sensitive with the learning rate and it failed in many instances to converge. On the other hand, Nadam, RMSProp

Table 4.8: Accuracy of GRUs and LSTMs with Different Parameters Settings

Hyper-parameters	Variants	GRUs Acc	LSTMs Acc
Optimizer	Nadam	94.26	93.08
	RMSProp	93.57	92.89
	SGD	71.88	79.20
	Adam	91.33	92.99
Batch_Size	1	55.72	54.55
	8	94.21	93.05
	32	94.26	93.08
	256	93.13	92.88
Activation Function	relu	94.26	93.08
	softmax	93.53	92.59
	sigmoid	94.06	92.64
	softplus	94.11	92.89
No. of. Rec Units	20	93.58	92.35
	40	93.87	92.84
	64	93.77	92.88
	128	94.26	93.08
	256	94.21	92.54

and Adam produced stable results of more than 91%. With respect to batch size, the mini-batch size of 1 produced poor accuracy. However, the algorithm achieved relatively good performance for batch size of 8 or more. Higher batch size value does not increase the performance of the models. Very big batch size of 256 seems to slightly decrease the conformance. The algorithm was also evaluated with different activation functions, including relu, sigmoid, softmax and softplus. The choice of activation function does not influence the performance of the algorithms as indicated by similar accuracies for both algorithms. Similarly, the number of recurrent units does not have any influence on their performance. Even though, the standard setting of 128 recurrent units appear to result in slightly better performance than other settings.

Table 4.9: Words have Significant Difference of Occurrence Likelihood between Classes.

Word	Critical	Uncritical	Difference	z-score	p-value
He	0.62	0.02	0.60	30.662	0.000
My	0.74	0.18	0.56	25.018	0.000
I	0.75	0.27	0.48	20.950	0.000
She	0.43	0.02	0.41	23.592	0.000
Him	0.39	0.01	0.38	23.367	0.000
We	0.35	0.18	0.17	8.494	0.000
Was	0.67	0.04	0.63	30.654	0.000
Is	0.70	0.37	0.33	14.469	0.000
Were	0.24	0.01	0.23	16.881	0.000
Will	0.36	0.16	0.20	10.177	0.000
Year	0.51	0.05	0.46	24.294	0.000
Abuse	0.46	0.12	0.34	16.910	0.000
Time	0.40	0.09	0.31	16.998	0.000
Life	0.39	0.08	0.31	17.055	0.000
Child	0.35	0.07	0.29	16.441	0.000
Friend	0.35	0.06	0.29	16.654	0.000
Husband	0.23	0.01	0.23	17.173	0.000
Night	0.24	0.01	0.22	16.634	0.000
Leave	0.24	0.03	0.21	15.332	0.000
Kill	0.22	0.01	0.21	16.112	0.000
Story	0.29	0.09	0.20	11.664	0.000
Love	0.29	0.12	0.17	9.820	0.000
Police	0.17	0.01	0.16	13.673	0.000
Woman	0.17	0.03	0.14	10.892	0.000
Survivor	0.24	0.11	0.13	7.897	0.000
Court	0.14	0.01	0.13	12.033	0.000
Control	0.14	0.01	0.12	11.676	0.000
Fear	0.14	0.01	0.12	11.287	0.000
Domestic	0.38	0.26	0.12	5.631	0.000
Victim	0.19	0.08	0.11	7.340	0.000
Violence	0.38	0.28	0.10	4.759	0.000

4.4.5 Semantic Coherence Analysis

This section first examines the data sets to identify important words that helps distinguish critical from uncritical posts. We computed the support of each word in their corresponding class, which reflects their likelihood of occurrence. The difference in the supports of each word between two classes are computed, and the words having highest differences are reported in Table 4.9. Z-test with $p - value \leq 0.05$ were performed to verify statistical significant of the difference.

In the DV corpus, stop-words such as *linguistic dimensions (I, she, he, my, him)* and *time oriented tenses (was, is, were)* are more associated with critical posts. We may understand that, when the victims or survivors post about their abusive experience, they use more past tense *(was, were)*. *"She, he"* notions are often used to refer the abusive partners. *"I, me, my"* are often used by the victims to express their sufferings. Example posts are: *(1) I am a survivor of DV and rape. I really need help right now. I was in a relationship with a man for 8 years. (2) He was cheating on her. When she confronted him, he hurts her. He is just an evil and greedy man.*

Besides, the words *"year*, *abuse*, *time"* are most likely to occurred in critical class than uncritical class with large differences. Those words usually appear in critical post, when victim made a post online to seek help from DVCS groups. The post content usually mentions about the victim was in an *abusive* relationship for the number of *years*, and when the last *time* the violence has happened. An example of such post is: *(He abused me for 5 years and each time he does something to scare me)*. Many posts mentioned about the context of the abusive incident, which is with the presence of their child/children, and sometimes, the child is also a victim. Thus, many critical posts contained the word *"child"*.

The word *"husband"* appears more in critical posts as male partner violence is predominant. Many posts mentioned that *husband* is abusive. Similarly, the word *"friend"* is used often in critical posts. Some posts mentioned that the victims called friend for help when the violence occurred, or sometimes male friend is mentioned as the abusive person in the posts. The words *"night"* describes the time of abusive incident, which is usually represented at night times. The abuse is either physical or sexual assault.

The words *"leave*, *love*, *control*, *fear, kill"* often occur in posts that represent the emotion of the victims and explaining the reason they want to stay or leave the relationship. An example of such posts is: *(I live in fear every night, that he will kill me and finally I decided to leave him)*. When the victim seeks legal support or guidance in critical situation, the words *"police* and *court"* usually occur in the posts.

We noticed that the words *"domestic*, *violence*, *abuse"* have high support in both critical class and uncritical class. Because, these terms *domestic violence* and *domestic abuse* are commonly used in difference context in relation to DV. They are often used in uncritical posts to create awareness messages such as *(lets spread the word on domestic violence against women, please share this page with your friends.)*

Although, the words presented in Table 4.9, highlighted some difference between critical and uncritical posts, solely relying on term frequency may not be effective in automatic classification of the posts. Because, some words are often used to gather and share similar meaning such as *domestic*, *violence*, *abuse*. The classification model should account for their semantic relationships rather than treating them as separate words as in the traditional features of Bag of Words and TF-IDF.

Fortunately, the word embedding features used in Deep Learning could be able to address this issue. Note that, each word is represented by a vector feature of 300-dimensions that captures its semantic meaning. Words with similar meaning would have similar vector features. In other words, vector features of similar words are highly correlated with each other. As a demonstration, we visualize the correlation between the embedding vector features for some sample

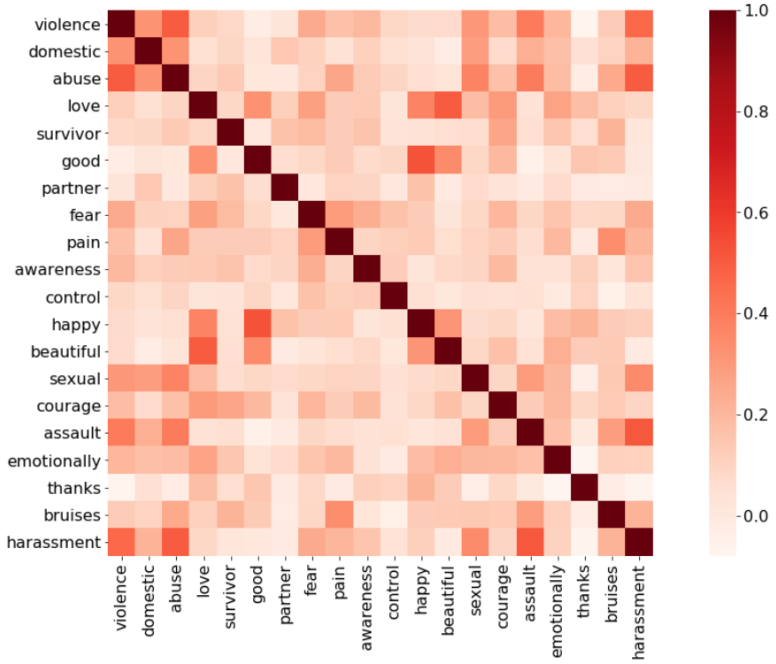


Figure 4.4: Correlation of sample words

words using a heat map in Figure 4.4. We can see that, there is a strong correlation between the word *abuse* and words *violence*, *harassment* or *assault*. There is a low correlation between words having difference meanings, such as *love* versus *assault*, *bruises* or *pain*. The word embedding features could be able to account for such relationship, which explain the higher performance of Deep Learning models in comparison with the traditional models.

4.5 Summary of Findings

In this chapter, we presented an approach for critical post identification using Deep Learning. The contributions of this work are: (1) A benchmark dataset was constructed from Facebook posts made by DV victims, with labels for critical and uncritical posts; (2) We evaluated the performance of the various Deep

Learning models in comparison with other traditional methods and with different parameter settings on DV critical post identification task. Due to the use of word embedding features, Deep Learning models (except for RNNs) achieved better performance than traditional models. The best setting for critical post identification of DV dataset is GloVe word embedding and GRUs model, with the Nadam optimizer and batch size of 32. Although, GRUs achieved the highest prediction rates in our experiments, other models CNNs, LSTMs and BLSTMs also achieved relatively high performance. Thus, Deep Learning models were demonstrated as promising to be adopted for developing practical solutions to identify the critical posts to support DV victims in critical needs. The analysis of the word occurrences also highlighted some context when and where DV take place. Future work, can consider classifying the posts into different DV context so that better detection of critical posts can be achieved and appropriate corresponding support can be provided to DV victims.

Despite the achieved results and findings, our work has several limitations. Namely, the data set used in the experiments was not at a big scale due to the labor-intensive job of manually labelling the posts. We currently recognized the critical post identification was mainly evaluated for posts from Facebook. Other Social Media platforms such as Twitter and Reddit can be considered in the future studies. Application for real-time critical post identification can be considered in the future so that instance support to DV victims can be provided. A novel algorithm for feature extraction or Deep Learning technique was not proposed in this work. Because, our primary focus is to evaluate the existing state of the art features and Deep Learning algorithms on the new research problem of critical and uncritical post identification. Nevertheless, the results and findings are valuable in guiding the future works on DV crisis identification.

CHAPTER 5

DEEP LEARNING FOR MULTI-CLASS IDENTIFICATION FROM DOMESTIC VIOLENCE ONLINE POSTS

DV is not only a major health and welfare issue, but also a violation of human rights. In recent years, DVCS groups active on Social Media have proven indispensable in the support services provision to victims and their families. In the deluge of online-generated content, the significant challenge arises for DVCS groups' to manually detect the critical situation in a timely manner. For instance, the reports of abuse or urgent financial help solicitation are typically obscured by a vast amount of awareness campaigns or prayers for the victims. The state-of-the-art Deep Learning models with the embeddings approach have already demonstrated superior results in online text classification tasks. The automatic content categorization would address the scalability issue and allow the DVCS groups to intervene instantly with the exact support needed. Given the problem identified, this chapter [1] aims to: *(i) construct the novel 'gold standard' dataset from Social Media with multi-class annotation; (ii) perform the extensive experiments with multiple Deep Learning architectures; (iii) train the domain-specific embeddings for performance improvement and knowledge discovery; (iv) produce the visualisations to facilitate models analysis and results interpretation.* Empirical evidence on a ground truth dataset has achieved an accuracy of up to 92% in classes prediction. The study validates an application of cutting edge technique to a real-world problem and proves beneficial to DVCS groups, health care practitioners and most of all - DV victims.

[1] *This chapter is based on the following article.*
Sudha Subramani, Sandra Michalska, Hua Wang, Yanchun Zhang, Haroon Shakeel, Jiahua Du "Deep Learning for Multi-Class Identification from Domestic Violence Online Posts". (Accepted: IEEE Access)

5.1 Introduction

In the previous chapter, the approach for *binary classification* of 'critical' versus 'non-critical' online posts using Deep Learning has been proposed and presented. In this chapter, we present the *multi-class posts categorisation*, providing the finer-grained insight into the violence prevalence and severity from online discourse. The online DVCS groups are promoted for safe advertisement of DV resources, awareness promotion about the need for compassion to victims, resource sharing, buddying between survivors, non-professional mentoring and fund-raising events to help victims and their families [64]. However, with the popularity of aforementioned initiatives, the online content generation has grown rapidly in scale. The unstructured and noisy character of such data has further added to an overall complexity of processing and utilising available information. The deluge of messages that are of personal nature in a form of mere opinions or empathetic thoughts along with general awareness promotions have greatly diminished the DVCS services efficacy to identify the critical situations and respond in a timely manner. Hence, in this chapter, we extended our previous work with multiple fine grained classes or more precise information categories, that supports DVCS groups. To further demonstrate the source of a problem identified, the exemplary messages and their corresponding labels have been presented in Table 5.1.

Information types of DV posts on Social Media: The automatic content categorisation allows the DVCS groups to efficiently handle the high-volume and high-velocity data, evaluate the nature of the problem, and respond almost instantly. After the posts analysis, 5 distinctive classes have been identified under the supervision of the experienced psychiatrist, active in family

Table 5.1: Examples of DV posts and the Corresponding Information Category

ID	DV Posts	Context	Label
P_1	*"After four years and an engagement, I realized for the first time, that I'd been in an abusive relationship with my fiance. I am deeply saddened by the idea that he does not know how to love me "*	*Shared by DV survivor*	*Personal Story*
P_2	*"Click here to support Lily And Nicole's Safety. Hello my name is Alice. I have decided to develop a Go fund Me donation account to help protect my children Lily and Nicole."*	*Seeking Financial help*	*Fund Raising*
P_3	*"Lets stop the violence. Know the signs."*	*Awareness promotion*	*Awareness*
P_4	*"Rest in Peace, Beautiful Angel."*	*Expressing empathy*	*Empathy*
P_5	*"Morning greetings. Enjoy the day."*	*Irrelevant to DV*	*General*

violence domain. The categories have been assesed based on their criticality and type of the support needed. In the post P_1, the victim seeks an emotional support from the community through sharing a personal experience with DV and the P_2 is an example of the financial aid solicitation. The 'Personal Story' as well as 'Fund Raising' categories have been given high priority as distinguishing two most common support needs (emotional vs financial) expressed in Social Media groups dedicated to DV. According to Evans et al., victims are experiencing long waiting times to access specialist healthcare services and those services are significantly underutilized [64] . As the 'Personal Story' category includes the abusive experience description, therefore its timely detection is of particular value not only for DVCS communities, but also for public health monitoring due to frequent inclusion of details of the violent incidents (physical, emotional etc.) and other health-related issues associated with the abusive experience (anxiety, depression etc.). As an example: *"I desperately need help. He physically assaulted me and threatened to kill me. I have spent the last*

10 months with depression, and PTSD". Considering the remaining classes, the 'early-intervention' and 'awareness-raise' were the violence prevention strategies proposed by WHO [162].

Thus, the posts P_3 & P_4 raise an awareness of DV and express an empathy with the victims. Finally, the post P_5 is classified as 'general' and does not assist DVCS groups in response coordination nor provide any additional insight into DV instances. The examples provided illustrate the need for an automatic posts classification in order to pro-actively support the potential DV victims and fulfill the WHO initiatives goals. Overall, an automatic thematic grouping is beneficial, even crucial given continuously growing DV community as a response to the wide prevalence of violence acts as well as the commonness of Social Media aid-seeking [43, 48, 123, 134].

The prioritization of content enables to efficiently reach out to potential victims with the exact support needed and in a timely manner, which is oftentimes a key determinant of the successful help endeavors. Nonetheless, the task proves extremely challenging given the rapidly growing Social Media data as well as lack of DV benchmark corpora for classification. The difficulty comes not only from the large volume and real-time posts inflow, but also the unstructured and highly noisy character of textual data.

As for the traditional text classification techniques, their performance accuracy rely heavily on the features extracted. Due to the unstructured format and informal character of Social Media data, manual feature engineering is considered tedious and ineffective. From the misspellings, through abbreviations, to synonyms, the automatic posts categorisation poses significant processing challenges in order to produce meaningful results. As an example, the alternative

term or shortened version such as 'Domestic Abuse' and 'DV' refer to the same concept of 'Domestic Violence'. Consequently, the basic search query for posts identification proves severly limiting.

Deep Learning has already proven successful in text classification tasks, outperforming the benchmark Machine Learning techniques [151] [20]. The most distinctive features are evaluated automatically during the model training process. To further improve the classification performance, the pre-trained embeddings are commonly incorporated into the model. The concept of embeddings is based on the assumption of terms semantic relationship, i.e. the pair *'assault'* and *'abuse'* will display closer distance in the vector space than the pair *'love'* and *'abuse'*. Still, the effectiveness of embeddings in classification tasks depend on the volume, quality, and the relevance to the domain knowledge of data used for their training. Thus, the domain-specific embeddings generation is getting increasing amount of attention among the researchers.

To the best of our knowledge, no prior study has performed multi-class DV posts identification nor evaluated Deep Learning algorithms against Machine Learning techniques with different features in DV context. The experiment results and analysis are beneficial to researchers, who are interested in carrying out further research in DV based on online Social Media data. Thus, the main objectives of this chapter are as follows:

- DV corpora creation with multi-class annotation ('gold standard');
- State-of-the-art Deep Learning models classification accuracies comparison;
- Superior performance of Deep Learning over Machine Learning empirical validation;

- Domain-specific embeddings construction from over 500k DV-related online posts;
- DV embeddings versus default embedding (GloVe) performance analysis;
- Knowledge discovery about the violence issue from Social Media.

Section 5.2 discusses about multi-class identification approach used in previous online text classification tasks and identified themes in different contexts. Section 5.3 covers methodology followed, namely: data collection from Social Media, 'gold standard' corpora construction, features extraction with pre-trained word embeddings, Deep Learning models specifications, and the performance metrics used. Section 5.4 details the experiment design and analysis, including Knowledge discovery from the pre-identified classes, features extraction and model training, the classification accuracies comparison, hyper-parameters explanation and evaluation, visualisation-supported performance and error analysis, and the DV-specific embeddings analysis. Section 5.5 concludes the results, highlights limitations and proposes future directions for the study.

5.2 Literature Survey

In the previous chapter (Chapter 4), the approach for binary classification of posts as either 'critical' or 'uncritical' has been proposed. In the current chapter, the focus was to further increase the effectiveness of DVCS groups in support of DV victims. Consequently, the more informative and finer-grained classes were derived under the supervision of qualified and active in family violence psychiatrist. Table 5.1 illustrates the exemplary posts for each class, and further

rationale for the classification is provided in the sub-section 5.3.2. As a result, this thesis addresses the two main challenges currently encountered by DVCS groups (1) irrelevant and noisy posts are filtered out giving importance to critical ones (previous chapter), (2) finer-grained categories identification (current chapter).

An approach followed was successfully adopted in prior studies on multi-class identification, where binary classification was performed as the primary step. As an example, the crisis-related Social Media posts have initially been classified into one of the two classes 'informative vs non-informative' [151]. In order to efficiently coordinate the crisis responses, further information categories were derived that related to affected individuals, donations, sympathy and support etc., and directed to different relief functions. Such approach facilitated the crisis situations management by the humanitarian organizations using state-of-the-art algorithms. Similarly, another study proposed automated classification approach for mental health-related posts identification from Social Media, followed by further categorization into the specific disorders such as bipolar, anxiety or depression, based on the underlying theme of the post [80]. Table 5.2 summarizes the related work in the areas of (i) disaster response and crisis management, and (ii) hate speech detection on Social Media. It further discusses about the various identified themes in multi class approach and the classifiers compared to identify the best performing model.

In order to accelerate disaster response and reduce human loss during natural disasters, Imran et al. [99] proposed automatic methods for the emergency responders to process the information of disaster affected communities in a timely manner. They categorized the posts into 5 categories, namely 'Caution &

advice, Information source, Donation, Causalities & damage, and Unknown', in terms of the context and usefulness of the posts. The Machine Learning algorithms were compared against various textual features such as unigrams, bigrams, part of speech tags and NB classifier achieved reasonably good performance. In the another similar study of crisis response, Nguyen et al. [151] categorized the Social Media posts based on information types such as 'Affected individuals, Donations and volunteering, Infrastructure & utilities, Sympathy & support, Other useful information, and Irrelevant' to assist humanitarian agencies in response coordination. Various algorithms such as SVM, NB, LR and CNNs were tested and the authors concluded CNNs model more suits the disaster response management, as the Deep Learning models has an advantage of automatic feature extraction, without any manual engineering process. To help emergency responders and public to react emergencies in a timely planner, Nicoli et al. [186] proposed 4 categories of messages as 'Emergency preparedness, Emergency response, Post emergency and recovery, and Engagement'. They trained classifiers such as CNNs, RNNs, and GRUs to automate the process for accurate classification of posts. The best performing classifier according to their experimental analysis is GRUs.

In the another domain of hate speech detection on Social Media, Thomas et al. [55] developed an approach for the separation of hate speech from offensive language. They categorized the posts into 3 types namely: 'Hate speech (racist and homophobic posts), Offensive (sexist posts) and Neither'. The LR classifier reached higher classification performance, when it was compared with NB, DT, RF and SVM. In the similar study of hate speech detection with the similar 3 classes, pikesh et al. [10] found, the classification performance led to the best accuracy, when the LSTMs combined with gradient boosted decision

trees (GBDT) than the other models such as LR, SVM, GBDT, CNNs, LSTMs. sreekanth at al. [127], classified an input post into 3 classes namely 'Overtly aggressive, Covertly aggressive, and Non-aggressive' for the problem of aggression detection on Social Media. They developed an majority voting based ensemble method by combining the 3 classifiers CNNs, BLSTMs, and LSTMs to predict the aggressive content on Social Media.

Despite the difference in domain, that could be either crisis response or abusive language detection, the promising results of classification models are increasingly observed in numerous real-time Social Media applications. These studies further prove that a careful design can yield an efficient system, leading the way for more sophisticated data analysis and interpretation systems. Nonetheless, all of the evaluated models demonstrate superior text classification performance, yielding comparable results. Still, the selection of the most optimal model is highly dependent on the application task as well as the hyper-parameters setting.

Inspired by the above-mentioned works, and to identify the best performing features in DV multi-class identification, the various sets of features extracted from the users' posts were compared. Further, the traditional Machine Learning algorithms with BoW, and TF-IDF features were comapred with the Deep Learning models that use word embeddings, to understand which model combination provide improved classification results. No attempts has been made to investigate the potential of deep learning in applications of multi-class identification of DV posts. This chapter aims to address the challenges in DV to predict the best performing model for multi-class identification problem.

Table 5.2: Example Systems in the Context of Disaster Response and Hate Speech Detection on Social Media

No.	Category	Social Media	No. of. categories	Multiple classes	Methods_compared	Accuracy
1	Disaster Response [99]	Twitter	5	"Caution & advice, Information source , Donation, Causalities & damage, and Unknown"	NB, SVM	NB
2	Crisis Response [151]	Twitter	6	"Affected individuals, Donations and volunteering, Infrastructure & utilities, Sympathy &support, Other useful information, and Irrelevant"	SVM, NB, LR, CNNs	CNNs
3	Emergency Response [186]	Facebook	4	"Emergency preparedness,Emergency response, Post emergency and recovery, and Engagement "	SVM, CNNs, GRUs	GRUs
4	Hate Speech Detection [55]	Twitter	3	"Hate speech, Offensive Language, and Neither"	LR, NB, DT, RF, SVM	LR
5	Hate Speech Detection [10]	Twitter	3	"Racist, Sexist, and Neither"	LR, SVM, GBDT, CNNs, LSTMs	LSTMs + GBDT
6	Aggression Detection [127]	Facebook	3	"Overtly aggressive, Covertly aggressive, and Non-aggressive"	CNNs, LSTMs, BLSTMs	Ensemble

5.3 Methodology

The section presents the proposed approach (Figure 5.1) for multi-class DV posts identification from Social Media. The methodology consists of the five steps, detailed in the following sub-sections.

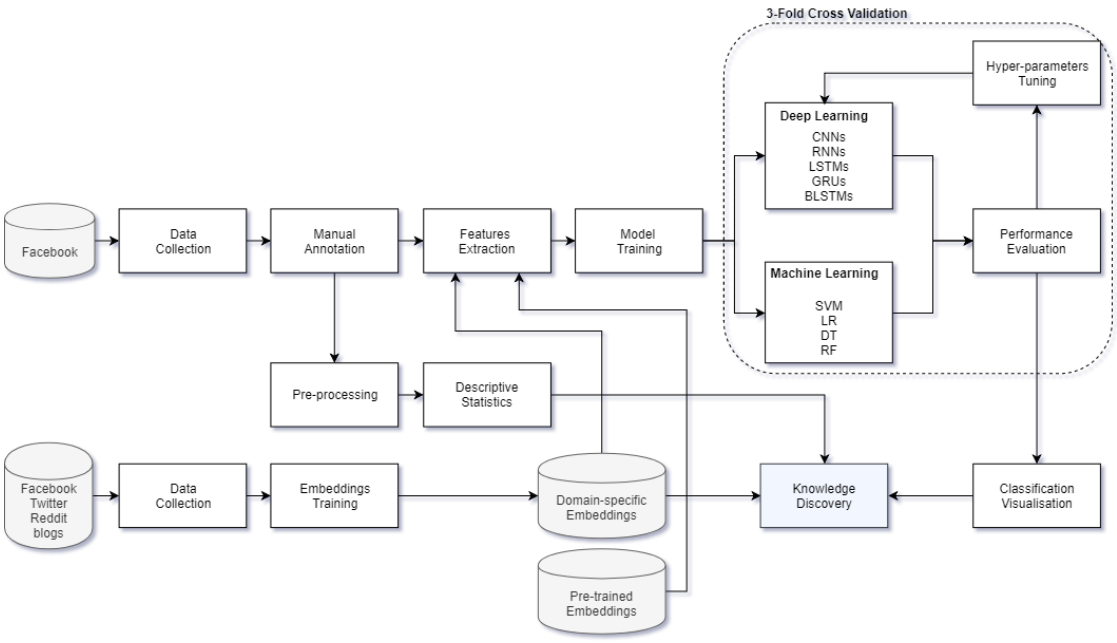


Figure 5.1: Architecture of Proposed Approach for Multi-class Identification

5.3.1 Data Extraction

Due to wide popularity and extensive engagement of sharers and supporters on DVCS Facebook group, the data was extracted from Facebook as the principal Social Media platform. Facebook is also ranked first among the top 15 social networking sites with around 2 billion users worldwide [191]. The posts were collected from pages that discuss the range of DV-related matters. The Facebook Graph API was used in the extraction process and the search terms were 'Domestic Violence' and 'Domestic Abuse'. A number of posts and comments of approximately $100,000$ was returned following the data collection from the 10 most active DV pages. The benefit of the Facebook Graph API [65] is that researchers can develop the applications to detect an information type of new posts in real-time, which can further enhance the DVCS groups efficacy. Considering the ethical concerns, the posts were collected solely from publicly available pages, and the identities of individuals included in the extracted dataset remained confidential.

5.3.2 Gold Standard Construction

In order to construct the 'gold standard' [2], the manual classification of data extracted was performed. Since human annotation is a time consuming process, the random $3,000$ posts were sampled. The instances containing only hyperlinks or images were excluded from further processing. The final benchmark corpora consisted of 1654 posts in total with a following breakdown between the categories: *Awareness* - 345, *Empathy* - 371, *Fund Raising* - 288, *Personal Story* -

[2] https://github.com/sudhasmani/DV_Dataset

352 and *General* - 298. The size of obtained dataset is considered moderate, given no previous work on multi-class DV online posts identification had previously been undertaken.

The posts were categorised as *Awareness*, *Empathy*, *Personal Story*, *Fund Raising* or *General* (Table 5.1). To further illustrate the annotation process, the exemplary messages, corresponding labels, and classification rationale have been presented in the following points:

- P_1 post as *Personal Story*: Emotional support from the community seeking through personal experience sharing (critical);
- P_2 post as *Fund Raising*: Financial assistance in the crisis moment solicitation (critical);
- P_3 post as *Awareness*: Awareness about the violence promotion (non-critical);
- P_4 post as *Empathy*: Empathy expression from community (non-critical);
- P_5 post as *General*: No additional insight into the DV problem (non-critical).

The annotation was performed by 2 research students under the supervision of a consultant psychiatrist with specialisation in DV field. Involvement of the domain expert was deemed necessary to ensure the credibity and usefulness of the 'gold standard' constructed. The Kappa coefficient was calculated to validate the inter-rater reliability as the most commonly used metric in similar type of studies [136]. The degree of agreement obtained was 0.81. In case of uncertainty, the final label was assigned following an advice of the expert.

The example of the borderline post is as follows: *"hi. my name is sarah. i am a domestic violence survivor with a brain injury from dv. i am mum to two beautiful children. she is 12 and he is nearly 3. i left my abusive former partner the day he attacked me with our 4 month old son in my arms. my son was traumatized in this violent physical attack. i have just recently finish a 5 year dream. i wrote a book. i would love to inspire other women and encourage them that we can all have our sacred loving self back to ourself and lead a normal happy life. i want to request you to put my latest book on your page. please support if you can"*. The post can be classified as *Personal Story* given that the victim shares her personal experience with DV, as well as implies the need of emotional support from DVCS community. On the other hand, the post can be labelled as *Awareness* provided that the problem had already been battled by the victim, who aims to promote her book inspiring other women in standing up against violence. Guided by the domain expert, the post was finally classified as *Personal Story* due to detailed depiction of the abusive relationship experience, and the potential for fine-grained knowledge extraction.

5.3.3 Feature Extraction

An important part of Deep Learning application to multi-class identification task involves the use of word embeddings as the features extraction. Words embeddings are considered the more expressive representation of text data, capturing the relationships between the terms. The vector representations of words are learnt in such a way that the similar concepts will be positioned nearby in the vector space. The unique characteristics of words embeddings such as automatic features extraction, semantic relationships retention and significant

dimensionality reduction overcome the drawbacks of the traditional features extraction such as sparsity and non-semantic representation. For instance, the terms 'depression' and 'anxiety' will be considered as distinctive features in the BoW model, which will only count their occurrence. The fact that both belong to the mental health condition category (thus being semantically related), would be ignored by the classier leading to decreased performance on the prediction task.

The two most common word embeddings that were trained on the large external corpus such as Google's Word2Vec [142] and Twitter's crawl of GloVe [181] have already shown promising results in various class prediction tasks. On the other hand, the domain-specific embeddings were also applied and validated, demonstrating the improved performance in text classification (crisis embeddings [150]) and named entity recognition (medical embeddings [251]) applications.

In order to evaluate the potential class prediction performance improvement using domain-specific embeddings, the DV embeddings have been constructed. The classification accuracy of Deep Learning models trained on pre-trained and DV-specific embeddings was then compared. The details of embeddings and experiments performed are as follows:

- **Pre-trained embeddings:** The two most popular embeddings have been used, namely Word2Vec and GloVe. The former has been trained on nearly 100 billion words from Google News, and covers 300 dimensional vectors for a vocabulary of 3 million words and phrases [142]. The latter has been trained on nearly 840 billion words from Twitter posts, and covers 300 dimensional vectors for a vocabulary set of 2.2 million words and phrases

[181]. Thus, for both feature sets, each word is represented by a vector of word embedding containing D = 300 dimensions.

- **Domain-specific embeddings:** The domain-specific embeddings [3] have been trained on the large corpus of DV-related discussions to differentiate from the generic news and tweets. The sources for data extraction included Facebook, Reddit, Blogs and Twitter. Only topic relevant posts were considered (e.g. victims support forums, abuse-dedicated groups etc.). In total, the corpus contained nearly 500k posts. The 50 and 300 embedding dimensions were used for training, given the relatively small size of the dataset in comparison with the pre-trained embeddings.

5.3.4 Model Development

The 5 Deep Learning models were adopted at this stage, namely:

- **CNNs:** The CNNs architecture used is described in-detail in [105]. In the first layer of the model, the most informative n-gram features are extracted, and the embeddings for each word are stored. Then, it passes through the pooling layer to produce feature vectors, and transforms the previous convolutional representation into a higher level of abstract view. Finally, the dense layer takes the combinations of produced feature vectors as input, and makes the prediction for the corresponding post.
- **RNNs:** The RNNs architecture used is described in-detail in [225]. RNNs handle a variable length sequence input by having loops called recurrent hidden state, which captures the information from previous states. At

[3] https://github.com/sudhasmani/DV_embeddings

each time stamp, it receives an input and updates the hidden state. The advantage of RNNs is that the hidden state integrates information over previous time stamps.

- **LSTMs, GRUs and BLSTMs:** LSTMs [84], GRUs [36] and BLSTMs [85] are improved versions of RNNs. The core idea behind LSTMs are memory units, which maintain historical information over time, and the non-linear gating units regulating the information flow. GRUs are basically the LSTMs with two gates, whereas LSTMs have got three gates. GRUs merge the *input* and *forget* gates into one unit called the *update gate*. BLSTMs consist of two LSTMs, which integrates the long periods of contextual information from both forward and backward directions at a specific time frame. This enables the hidden state to store both the historical, and the future information. Thus, LSTMs, GRUs and BLSTMs are considered the state-of-the-art semantic composition models for text classification tasks, which learn long-term dependencies between the words in a sequence without storing the redundant information.

5.3.5 Performance Evaluation

The Precision, Recall, F-Measure and Accuracy are selected as evaluation metrics for the classifier. These metrics have been used widely in previous studies to examine models performance [10, 150].

Also, the k-fold cross-validation was applied to assure the robustness of the validation and to prevent overfitting and the potential selection bias [31, 115]. The collected dataset was randomly divided into k partitions, where one par-

tition was reserved as the testing set, while the others were combined into the training set. The procedure was repeated k times for different testing sets. The results were averaged to produce the final performance metric.

5.4 Experiment Design and Analysis

In this section, the automatic classification experiments for categories identification from DV posts are discussed in detail. Several steps were performed to evaluate the performance of the introduced approach using Deep Learning. These include:

(A) *Descriptive Statistics:* The insights about the corpus characteristics such as number of posts in each class, the maximum and average words count in each class, before and after pre-processing. Also, the most frequent words in each class were produced for qualitative analysis.

(B) *Model Training:* The detailed steps for model training are presented including features extraction approaches (e.g. Word2Vec and GloVe), the rationale behind their application as well as the settings selection. The models training procedure is described for both Deep Learning and Machine Learning techniques.

(C) *Accuracy Evaluation:* The performance of the 5 Deep Learning models, namely CNNs, RNNs, LSTMs, GRUs and BLSTMs on the constructed benchmark data set was evaluated. Additional experiments with Machine Learning approaches, namely SVM, RF, LR and DT were conducted for comparison purposes. The most commonly used validation metrics, i.e. Precision, Recall, F-Measure and Accuracy were calculated.

(D) *Hyper-Parameters Evaluation:* Given an influence of the associated hyper-parameters on the performance of classifiers, the number of experiments with various settings were perfomed. The parameters optimised included the pre-trained word embeddings, optimizer type, dropout rate, number of recurrent units, and number of LSTM memory units, or convolution filters. As training and tuning a neural network can be time consuming, the selected parameters followed the study by Reimers et al. [196].

(E) *Models Visualisations:* The scatter plots and confusion matrices that visually depict the various Deep Learning architectures performance were produced. The graphical representation not only allows to obtain an instant overview of the similarities between the classes, but also to identify the main sources of misclassifications. As a result, the outputs interpretation is facilitated, and the potential errors better understood.

(F) *Domain-specific Embeddings Analysis:* The experiments were conducted to test our hypothesis of DV-specific embeddings over the generic pre-trained embeddings performance improvement. The analysis covered (i) an impact of the proposed embeddings on classification accuracy, and (ii) the insights and knowledge discovery about DV from the embeddings generated.

5.4.1 Descriptive Statistics

The descriptive statistics has been performed on the dataset with various pre-processing steps applied for comparative purposes, namely: (i) No pre-processing, (ii) Stopwords removal, and (iii) Stemming only. The total number of words in each class was calculated along with the average and maximum

number of words per post, also in each class. Finally, the most frequently occurred terms were extracted for finer-grained insight into the nature of each class.

From Table 5.3, we can observe that the total number of words was reduced significantly after the stopwords removal. This indicates a considerable proportion of generic vocabulary in the posts collected. Also, the discrepancy was noticed after stemming application as the words count increased for certain classes, e.g. the word 'F.B.I' was transformed into 'f', 'b', 'i'. Furthermore, the stemming procedure proved meaningless from the interpretation and knowledge discovery point of view, which can be illustrated with the example of terms such as 'domestic', 'abuse' and 'peace', converted into 'domest', 'abus' and 'peac'.

The most notable difference between the classes was observed in the total number of words, and the related average post length. The total number of words in *Personal Story* category accounted for 75% of the total number of words in all classes. In contrast, the *Empathy* category comprised only 3% of the total words count. Since victims share their personal experience, the lengthier posts in *Personal Story* category were expected, demonstrating the potential for deeper knowledge mining and discovery. On the other hand, brief *Empathy* posts such as 'Rest in Peace, we miss you beautiful angel' prove little informative with regard to the problem of violence, thus are considered non-critical in DVCS support services efforts.

Overall, with the presence of stopwords, the most frequent words include mainly prepositions, pronouns and articles, which apply to all of the classes. After stopwords removal, the valuable and interesting insights about the specifics

of each class emerge. The findings from the most common words are discussed with respect to the each class:

- *Personal Story:* The terms relating to the time/length of the abusive incidence such as 'years' or 'time' have been observed. Also, the dominance of the 1^{st} and 3^{rd} person pronouns i.e. 'me', 'my', 'he', 'she' or 'her' is characteristic for *Personal Story* category. This can be explained by the self-expressive nature of such posts, as well as the indication of the perpetrators e.g. 'He abused me for 10 years'. As demonstrated, even the stopwords add value in classes differentiation.
- *Fund Raising:* The terms 'support', 'click' and 'help' are the most prevalent, as expected. Additionally, the most common support recipients i.e. 'children', 'mother' are highlighted. As an example - 'please help me, please support my children'. Similar to Personal Story category, the 1^{st} and 3^{rd} person pronouns have been widely observed in this class as well.
- *Awareness:* Similarly to *Fund Raising* category, the most frequent words include 'please', 'share', 'support', 'like', 'love', 'awareness' as expected. Such terms do not provide additional insight into DV problem, therefore their classification as non-critical.
- *Empathy:* The most sympathetic words among all the classes such as 'sad', 'beautiful', 'heartbreaking', 'tragic' and 'love' have been observed. The main intention of the posts in the *Empathy* class is to show compassion to the victims, therefore their non-critical nature from DVCS perspective.
- *General:* The class dominated by generic, and mostly non-abuse related terms, including 'like', 'love', 'everyone', 'favorite', 'hey' or 'answer'.

Table 5.3: Exploratory Data Analysis of Multiple Classes

Pre-processing steps	Words count	Personal Story	Fund Raising	Awareness	Empathy	General
No Pre-processing	Total No of words	125631	19280	23346	5956	7269
	Max words count of posts	4310	143	1659	303	131
	Avg words count of posts	356	66	67	16	24
	Most Common Words	I, to, and, the, my, a, he, was, of, me, in, that, for, it, with, is, her, she, this, on.	to, I, and, a, my, the, of, by, support, here, click, in, is, her, for, was, domestic, she, help, with.	to, and, the, you, a, of, I, is, for, in, your, that, are, this, domestic, with, we, it, have, be	in, so, to, rest, and, peace, the, of, this, beautiful, all, is, sad, that, these, you, I, a, are, my.	I, a, to, the, my, you, and, is, on, of, have, for, that, in, with, what, it, so, your.
Stop words removal	Total No of words	57237	10224	13033	3074	3781
	Max words count of posts	2014	65	898	156	61
	Avg words count of posts	162	35	37	8	12
	Most Common Words	Time, one, back, life, never, know, like, help, abuse, years, violence, still, could, go, domestic, going.	Support, click, domestic, help, violence, children, mother, organized, abuse, years, name, family, abusive, need, life, home.	Domestic, violence, please, share, help, abuse, life, support, like, page, know, love, awareness, women, people.	.rest, peace, sad, beautiful, heartbreaking, lives, women, violence, souls, heart, angels, love, tragic, lost, breaks.	Like, love, everyone, good, get, day, want, hey, going, something, today, go, favorite, one, answer.
Stemming Applied	Total No of words	128511	19545	24596	5908	7193
	Max words count of posts	4347	151	1712	299	131
	Avg words count of posts	365	67	71	15	24
	Most Common Words	I, to, and, the, my, a, he, me, was, of, in, it, that, for, her, with, t, is, him, she.	To, I, and, a, my, the, of, by, support, here, click, in, her, is, for, was, domest, help, she, violence	To, and, the, you, a, of, I, is, in, for, your, that, this, be, domest, are, violence, with, abus.	In, so, to, rest, peac, and, the, sad, this, of, all, beauty, is, I, that, you, these, a, it, mani	I, a, to, the, my, you, s, and, it, is, have, of, for, that, what, in, t, all, with, do.

As presented, the manual features extraction is found interesting in terms of the potential insights and knowledge generation. Nonetheless, the approach proves less effective in classification task, with more time and effort required. In the following sub-sections, the extensive experiments will be demonstrated to analyse and compare the performance of traditional and advanced feature engineering methods.

5.4.2 Model Training

In order to examine the robustness of the classifiers, the features for Deep Learning models were extracted using the 2 main word embeddings, namely Word2Vec and GloVe.The first layer of the model is the embedding layer that computes the index mapping for all the words in the vocabulary and convert them into dense vectors of fixed size by parsing the pre-trained embedding. The subsequent layers contain 128 memory cells, which is the number popularly used in previous applications [196]. Additionally, the models were trained up to 50 epochs and implemented using Keras [37].

In Deep Learning, the pre-processing is not carried out as models process the sequence of words in the order they appear. Stopwords might hold valuable information that could be leveraged. Words are preserved in their original form without stemming as they can represent different context (e.g. the words 'abusive', 'abuser', 'abuse' are context dependent). Also, Nadam optimizer was used for Deep Learning models. Batch size was set to 32 posts as the dataset size was moderate. Relu activation function was used and recurrent units were set to 128. Dropout is an effective technique to regularize the model and combat overfitting [72, 219]. Accordingly, the dropout rate was set to 0.2 [196].

In terms of the traditional Machine Learning techniques, the most common feature models in text classification tasks i.e. TF-IDF and BoW were adopted. In order to overcome the limitation of our previous work [223], i.e. simple versus strong features and models comparison, the comprehensive experimentation with all of the potential 'feature-model' combinations was considered. For the evaluation purposes, the default parameters settings from python scikit-learn package were selected.

5.4.3 Accuracy Comparison

The dataset was partitioned into training and testing sets, following 3-fold stratified cross-validation approach, as used in previous studies [185] [172] [93]. The 3 pre-processing cases for traditional classifiers were selected:

(a) stopwords removal only;

(b) stemming only;

(c) both stopwords removal and stemming.

The Machine Learning models performance heavily depends on the pre-processing procedures undertaken. The results indicate that the traditional classifiers have achieved the highest performance with stemming only (b) (i.e. with stopwords retained). In the context of DV multi-class identification, some stopwords could be helpful in classes distinction (e.g. *Personaly Story* due to the large proportion of 1^{st} and 3^{rd} person pronouns).

The results also assist in identification of the most optimal case for comparison with the 5 Deep Learning architectures. Due to the space constraints, only the evaluation outputs for the highest-performance setting (b) for Machine Learning technique are shown. Evaluation metrics such as Precision, Recall, F-Measure and Accuracy were computed and are presented in Table 5.4.

Overall, Deep Learning models with GloVe embedding, which proved superior to Word2Vec, achieved improved performance over the traditional Machine Learning classifiers (except for RNNs), as indicated by the higher evaluation metrics outputs. In terms of the lowest score of RNNs, it can be attributed to the problem of vanishing gradients [14]. Given a long sequence, information of initial sequence fades away as the new sequences are fed into the networks of RNNs. Nevertheless, such limitation of RNNs seems to be overcome by its later versions, namely LSTMs, GRUs and BLSTMs. The successive versions can capture long-term dependencies efficiently, which is suitable for dealing with sequential textual data.

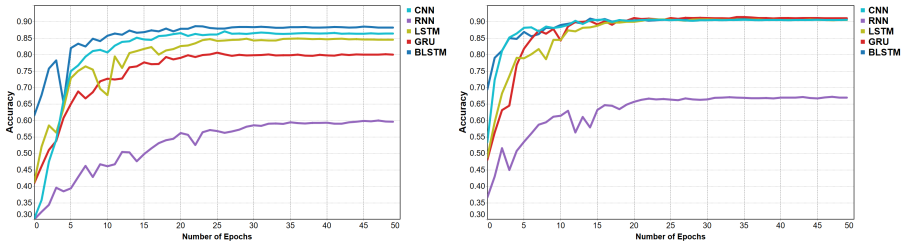
With GloVe embedding, GRUs and BLSTMs performed the highest with scores of 91.78% and 91.29%, respectively. RNNs achieved the lowest accuracy

of 67.65% among all 5 Deep Learning classifiers. With Word2Vec embedding, BLSTMs scored the highest accuracy of 89.12%. Still, its overall performance is lower than both the GloVe embedding and the selected Machine Learning classifiers, such as SVM and LR with TF-IDF setting.

Table 5.4: Evaluation Metrics of Classification Models

Model	Feature-Set	Precision	Recall	F-Measure	Accuracy
CNNs	Word2Vec	87.66	87.33	87.50	87.30
RNNs	Word2Vec	62.33	60.00	61.17	60.03
LSTMs	Word2Vec	85.33	85.33	85.33	85.25
GRUs	Word2Vec	81.66	81.00	81.33	81.14
BLSTMs	Word2Vec	89.33	89.00	89.17	89.12
CNNs	GloVe	91.33	91.00	91.17	90.93
RNNs	GloVe	69.33	67.66	68.50	67.65
LSTMs	GloVe	91.00	91.00	91.00	90.99
GRUs	**GloVe**	**91.66**	**91.66**	**91.66**	**91.78**
BLSTMs	GloVe	91.66	91.33	91.50	91.29
SVM	Word2Vec	88.98	88.26	88.62	88.36
LR	Word2Vec	88.50	87.49	87.99	87.64
DT	Word2Vec	63.45	61.78	62.60	62.55
RF	Word2Vec	77.12	76.63	76.88	77.09
SVM	Glove	88.10	87.39	87.74	87.45
LR	Glove	86.99	86.16	86.57	86.36
DT	Glove	64.23	62.41	63.30	62.91
RF	Glove	77.79	77.27	77.53	77.82
SVM	TF-IDF	91.00	91.00	91.00	90.81
LR	TF-IDF	91.00	90.33	90.67	90.45
DT	TF-IDF	82.33	82.33	82.33	82.29
RF	TF-IDF	86.00	84.33	85.17	84.40
SVM	BoW	88.00	86.66	87.33	86.58
LR	BoW	89.00	88.33	88.67	88.21
DT	BoW	83.66	82.66	83.16	82.77
RF	BoW	77.33	74.66	76.00	75.09

Table 5.4 results further demonstrate that Machine Learning models such as SVM and LR obtained higher accuracy with TF-IDF features. The Machine Learning classifiers are well suited for high dimensional and spare features vectors. It is obvious from the results that such classifiers are not suitable for dense vector representations with 300 dimensions. As the word embeddings are superior to traditional features, the advanced Deep Learning models can effectively



(a) Word2Vec (b) GloVe

Figure 5.2: Accuracy of Deep Learning Models at Different Epochs.

use the dense representation of words embeddings.

5.4.4 Hyper-parameters Evaluation

The performance of Deep Learning models was evaluated with respect to the training epochs. Ideally, more training epochs would result in the well-trained and stable models. However, Deep Learning often takes a long time to run. Setting high number of training epochs results in significant and unnecessary costs incurred. Figure 5.2 shows the accuracy of Deep Learning models using both Word2Vec and GloVe word embeddings against various training epochs. The models appear to converge faster with GloVe than Word2Vec features set. With Word2Vec embedding (Figure 5.2a), the accuracy of Deep Learning models fluctuated at the beginning and became stable after 30 epochs on average. With GloVe embedding (Figure 5.2b), the majority of the models reached stability after 20 to 25 epochs, except RNNs. Thus, the Deep Learning models arrived at the optimal accuracy and consistency in learning rate with the minimum training epochs using GLoVe embedding.

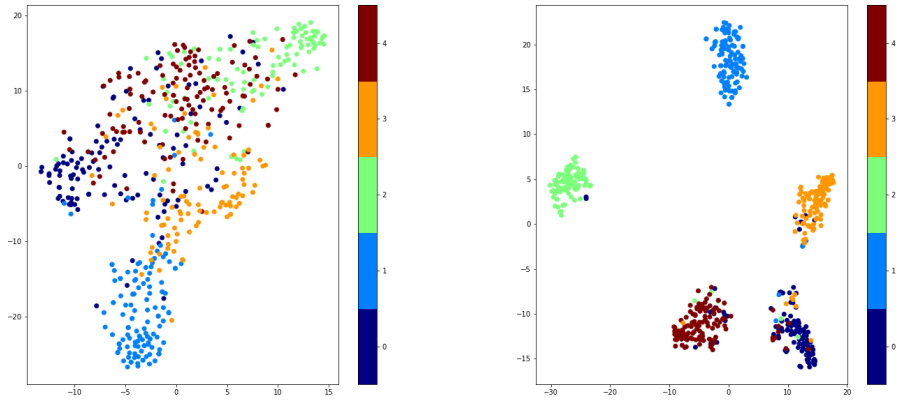
Next, the various hyper-parameters settings such as optimizer, batch size, number of recurrent units and activation function were evaluated on GRUs and

Table 5.5: Accuracy of GRUs and BLSTMs with Different Parameters Settings

Hyper-parameters	Variants	GRUs Acc	BLSTMs Acc
Optimizer	**Nadam**	**91.78**	**91.29**
	RMSProp	91.95	91.65
	SGD	42.85	51.42
	Adam	88.03	88.45
Batch_Size	**32**	**91.78**	**91.29**
	64	90.99	91.11
	256	69.22	90.15
Activation Function	**relu**	**91.78**	**91.29**
	softmax	91.11	91.12
	sigmoid	91.90	91.29
No. of. Rec Units	20	83.67	90.45
	40	89.18	90.51
	64	90.20	91.11
	128	**91.78**	**91.29**
	256	91.66	91.17

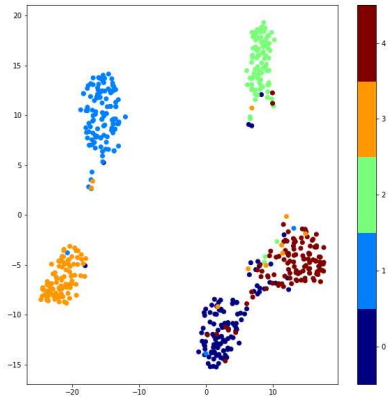
Note: The model training parameters, defined in sub-section 5.4.2 are highlighted in the table.

BLSTMs models, as their accuracy scores were the highest. The 3-fold cross validation was performed and the outputs are presented in Table 5.5. Among the optimizers, SGD is quite sensitive with regard to the learning rate and it failed in many instances to converge. Though Nadam, and RMSProp produced stable results of more than 91%, the computation time of RMSProp is much higher. With respect to the batch size, the algorithm achieved relatively good performance with the batch size of 32. Higher batch size value does not increase the performance of the models and the large size of 256 seems to decrease the conformance. The algorithm was also evaluated with different activation functions, including relu, sigmoid, and softmax. The choice of activation function does not influence the performance of the algorithms as indicated by similar accuracies for both algorithms. Similarly, the number of recurrent units does not have any influence on their performance. Even though, the standard setting of 128 recurrent units appear to result in slightly better performance than the other settings.



(a) RNNs

(b) BLSTMs

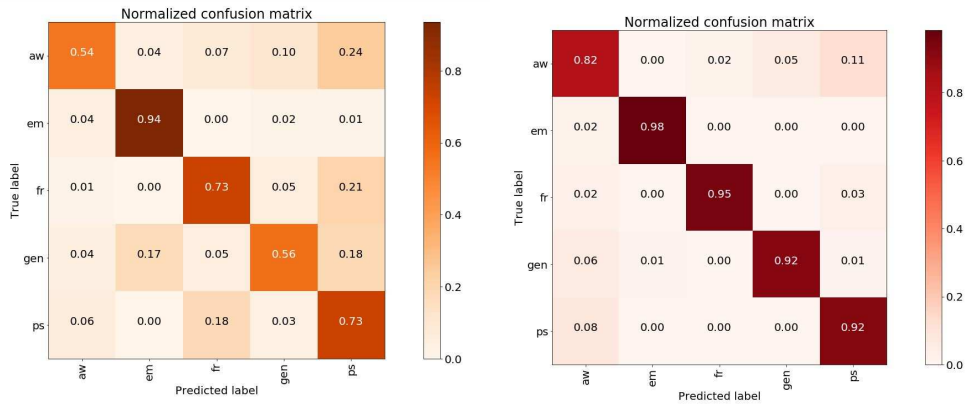


(c) GRUs

Figure 5.3: Visualization of Various Information Categories of DV Dataset using t-SNE w.r.t. GloVe Embedding. *(0-Awareness; 1-Empathy; 2-Fund Raising; 3-General; 4-Personal Story)*

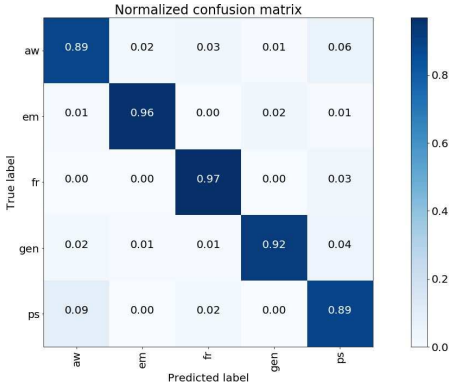
5.4.5 Models Visualisations

The models visualizations provide graphical insight into the classification of DV posts among various Deep Learning architectures. The dimensionality reduction technique t-SNE based on GloVe embedding was applied in order to plot the similarity between the categories. The highest (GRUs and BLSTMs) and



(a) RNNs

(b) BLSTMs



(c) GRUs

Figure 5.4: Confusion Matrix of Deep Learning Models w.r.t. GloVe Embedding. *(aw-Awareness; em-Empathy; fr-Fund Raising; gen-General; ps-Personal Story)*

lowest (RNNs) performing models were presented for comparison. The natural clustering between DV posts from their respective groups can be observed on scatter plots in Figure 5.3. From the analysis, the following conclusions can be drawn:

- The posts separation by RNNs (Figure 5.3a) model did not produce clear distinction between the classes. Additionally, the further overlap for *Fund Raising* and *General* categories occurred. The only clearly segmented group by RNNs model was *Empathy*. It is due to the specific characteristics of posts from that group such as very short and mostly repetitive phrases,

distinct from the remaining classes, e.g. 'Heartbreaking', 'Rest Peacefully Beautiful Souls'.

- The posts were segmented clearly into their belonging classes by BLSTMs model (Figure 5.3b). Minor misclassifications occurred between *Awareness* and *Personal Story* categories due to their content similarity. As an example, personal experience sharing motivated by the awareness raise to prevent future DV instances, e.g. 'I strive to raise awareness for this as even if it can make one person realise they are strong enough to get out, it's worth it.' (*Personal Story*)
- The posts were distinguished mostly correctly by GRUs model (Figure 5.3c) as well. Nonetheless, the overlap between *Awareness* and *Personal Story* classes was observed in 2D space resulting in higher number of misclassifications within those groups.

To further validate the findings and quantify the classification accuracy between categories, the confusion matrices were produced for the same models (RNNs, BLSTMs, and GRUs) (Figure 5.4). The 3-fold cross-validation approach was adopted. To reduce the potential interpretation bias, the fold with the highest score for all 3 models was considered. Similarly to the outputs generated by the scatter plots, the BLSTMs and GRUs proved the highest accuracy among all groups, i.e. 92% and 89% posts were classified correctly as *Personal Story* by BLSTMs (Figure 5.4a) and GRUs (Figure 5.4c), respectively. In terms of RNNs (Figure 5.4a), the group with the most misclassifications proved to be *Awareness*, where 24% of the posts fell under the *Personal Story* category. It did not perform well for *Fund Raising* and *General* neither, confusing them with *Personal Story* group as well. Overall, both the GRUs and BLSTMs maintained strong performance for various classes and therefore complement each other.

Table 5.6: Accuracy of Deep Learning Models with GloVe and DV Embeddings of 50 & 300 Dimensions.

Embedding Dimensions	Deep Learning Models	GloVe Embedding	DV Embedding
50 Dimesions	CNNs	90.27	**91.16**
	RNNs	53.87	**62.39**
	LSTMs	89.60	**90.69**
	GRUs	90.63	**91.35**
	BILSTMs	89.30	**91.23**
300 Dimesions	CNNs	**90.93**	89.96
	RNNs	**67.65**	61.27
	LSTMs	**90.99**	88.56
	GRUs	**91.78**	90.14
	BILSTMs	**91.29**	90.55

To further obtain the deeper insights into how the Deep Learning technique benefits the classification performance with respect to more fine grained content categorization, some of the posts from test dataset were sampled in Table 5.7. The labels predicted by the Deep Learning models (RNNs, BLSTMs, and GRUs) were compared with gold standard labels. From the given posts P_1 to P_9 (Table 5.7), it is clear that the RNNs wasn't able to correctly predict the classes and couldn't differentiate the underlying context. In the case of BLSTMs and GRUs, most of the posts were clearly segmented into their respective classes with the higher prediction probability, as previously discussed. However, the minor misclassifications occurred between *Awareness* and *Personal Story* classes. For instance, the actual class of the posts P_8 and P_9 were labelled as *Personal Story*. However, the BLSTMs and GRUs predicted the class of the posts P_8 and P_9 as *Awareness* respectively. This is because, though the posts P_8 and P_9 begin with personal sharing of abusive experience, the stories finally motivated by the awareness raise to prevent further DV instances and social support message.

Table 5.7: Sample Posts of Test Dataset and the Corresponding Prediction Results

ID	Posts	Actual_Class	RNNs	BLSTMs	GRUs
P_1	*"Negative emotions like hatred destroy our peace of mind. Guide your self to find peace and healing within you and other people. Hope we become better people to love and share the best in us. Feel free to contact me for any spiritual guided guidance and reading in any challenge bless"*	**aw**	ps (0.587)	**aw** (0.985)	**aw** (0.982)
P_2	*"We need to know what they look like how to respond how to help a colleague and how to create an environment where someone feels comfortable enough to go to someone confidentially and safely about needing support"*	**aw**	ps (0.732)	**aw** (0.989)	**aw** (0.996)
P_3	*"This is so true. They pull you in with their dream come true kind of world and then the hell begins"*	**aw**	em (0.520)	**aw** (0.865)	**aw** (0.988)
P_4	*"Here is a domestic violence awareness project done in michigan. we would love for this project to be shared with the world in hopes of speaking to others. These brave woman stepped out of their comfort zones in regards to helping each other heal as well as hoping it would reach those that are in abusive relationships to see they to can have the strength to leave. Thank you so much for your time and have a wonderful day. @username added 49 new photos to the album body shaming domestic violence awareness project we are now in the month of october and being that october is domestic violence awareness month. we gave woman around michigan the challenge of speaking out and having a voice.Please help share our stories and hopefully give others the courage to stand up and say no to domestic violence body shaming no more."*	**aw**	ps (0.442)	ps (0.896)	**aw** (0.490)
P_5	*"Rest in peace @username a friend who died by the hand of her ex boyfriend 20 years ago"*	**em**	aw (0.300)	**em** (0.992)	**em** (0.999)
P_6	*"I don't know if I want to put a sad emoticon or a angry one. far too many."*	**em**	gen (0.650)	**em** (0.955)	**em** (0.983)
P_7	*"Click here to support she broke her silence. she's a hero by @username. @username story, I believe everyone has the power to be strong enough to report rape to the police. After two weeks of suffering in silence on your own scared. Help reward survivors of rape especially after they broke their silence. Breaking your silence makes you a complete hero."*	**fr**	ps (0.535)	**fr** (0.978)	ps (0.717)
P_8	*"I am mum to two beautiful children s 12 and nearly 3 in june 2014. I left my abusive former partner the day he attacked me with our 4 month old son in my arms. My son was traumatized in this violent physical attack. I knew that day that not only was I in danger, but the lives of two innocent children were in danger as well that day the day i left for good. I promised my children i would keep them safe always and have prioritized their safety over and above anything else during the course of my four year relationship. I was abused physically verbally emotionally socially and financially..... I speak for every mother who has fled to safety with her children. Now our children are our future, they are our world. we will protect them from all forms of violence as is their basic human right no child can reach their full potential living with abuse neglect and family violence or the everlasting effects. I believe something needs to be done our children of domestic violence need help."*	**ps**	fr (0.476)	aw (0.904)	**ps** (0.988)
P_9	*"October is national domestic violence awareness month. Today marks the 16 month anniversary of my attack. I am a survivor of domestic violence. my attacker was charged and found guilty of a third degree felony strangulation asphyxiation charge against me permanent felony record but by the grace of god i escaped and survived hours of being held hostage in a hotel room. The 3 separate asphyxiation attacks and my 6 foot 4 inch attacker not even realizing the abuse that had been happening for the 5 months leading up to it, why i don't know. I was trying to find the good to empathize to heal my attacker..... Here are a multitude of wonderful resources and help groups for those in need for me the houston area women s shelter provided great support take action be active not passive. Forgiveness hope is the anchor of our soul. October is national domestic violence awareness month."*	**ps**	aw (0.642)	**ps** (0.758)	aw (0.845)

Table 5.8: Example Words and their Top Similar Words from User Posts using DV and GloVe Embeddings

DV related words	Learnt by DV embeddings	Learnt by GloVe embeddings
Abuser	attacker, assailant, perpetrator, spouse, affair, rapist, narcissist, ex, ultimatum, partner, victim, perp, husband, aggressor, predator	abusers, addict, pedophile, offender, rapist, addiction, molestation, abuse, abusing, psychopath, sex, alcholic, pornography, psychotic, prostitution
Abuse	violence, assault, degradation, cruelty, trauma, violenc, abusers, dv, violance, voilence, victimization, scarring, harassment, homicide, coercion	sexual, harassment, abuses, sex, criminal, rape, cases, torture, crime, neglect, discrimination, alleged, allegations, serious, charges
Domesticabuse	coercivecontrol, domesticviolence, wearpurpleinoctober, getout, dvam, silencetheviolence, childsexualabuse, cocoawareness, ipledgebecause, seedv, stopallviolence, stopabuse, domesticabuseawareness, stopthesilence, lovedonthate, healdonthurt, gbv, takeastand, embracingmyself, financialabuse, gethelp, seekhelp, reachout, childabuse, selfdefense, respectisvital, stayinfocused, voiceshavepower,vaw, domesticviolenceawareness, healingandrecovery, growinfight, gutsoverfear	prey, spinach, hides, realises, carton, leaves, feed, destined, parasite, discovers, unwittingly, infected, sickly, toxic, likelihood, lighttouches, venom, nutrients, realizes, virus, predators, grows, poisoning, berries, milk, pregnant, injecting, drained, returns, overdose, humans, liver, feeding, attractive.

Note: The words in the table also include misspellings and abbreviations, as they are more common in the users' postings on Social Media.

5.4.6 Domain-Specific Embeddings Analysis

The domain-specific embeddings for DV were constructed and evaluated in terms of (i) impact on the classification performance in Deep Learning models, (ii) useful insight generation and knowledge discovery about the DV. The comparative analysis with GloVe as a benchamark has been performed, and the moderating effect of different embedding dimensions was evaluated.

In terms of the average classification accuracy, the difference between GloVe and DV embeddings was negligible, which may not offset the time and cost involved in the domain-specific embeddings generation. The performance also varied across different dimensions levels, i.e. DV embeddings scored higher with the size of 50, whereas GloVe proved superior with the size of 300. From

computational efficiency perspective, lower number of dimensions is preferable, given the reduced training time. Table 5.6 presents the accuracies obtained for the CNNs, RNNs, LSTMs, GRUs and BILSTMs models, trained on both GloVe and DV embeddings, with the dimensions set to 50 and 300.

The other aspect of analysis involves the potential for insights generation and knowledge discovery from both GloVe, and the domain-specific embeddings. DV embeddings were trained on the content crawled from online forums dedicated to the violence victims support. Hence, it is expected that the results obtained from domain-specific embeddings will be more meaningful and valuable, than the outputs from GloVe, trained on the general Twitter corpora. To test the assumption made, the most similar words to the 3 DV-related terms were extracted from both GloVe and DV embeddings (Table 5.8). The similary was evaluated by the standard measure of cosine distance in the vector space. The 3 DV-related terms, such as 'abuser', 'abuse' and 'domesticabuse' were selected based on their common ocurrence in the user posts.

For instance, the sample words associated with 'abuser' generated by DV embeddings allow for deeper insight into, in this case, the types of victimisers, i.e. 'ex', 'partner', 'husband'. On the other hand, the generic nature of terms produced by GloVe returns more general terms.

As another example, the word 'abuse' represents the nature of abuse in the domain context. The DV-embeddings return the similar meanings of abuse term such as 'violence', 'assault', 'harrasment', and 'coercion'. The domain-specific embeddings proved superior in the identification of the commonly used abbreviations, such as 'dv' (Domestic Violence), and mis-spellings ('violenc', 'voilence').

Similar to 'abuse', the phrase 'domesticabuse' and its related terms returned by the domain-specific embeddings allow for deeper insight into, in this case, (i) the other types of abuse (e.g. 'coercive control', 'child sexual abuse', 'financial abuse', 'child abuse', etc.,), (ii) awareness messages (e.g. 'wear purple in october', 'get out', 'stop the silence', 'love dont hate', 'heal dont hurt', 'embracing myself', etc.,) and (iii) some common abbreviations such as 'dvam' - domestic violence awareness month and 'gbv' - gender based violence. In contrast, GloVe returns mostly irrelevant to DV terms such as 'prey', 'spinach', 'nutrients' or 'berries'. Thus the GloVe embeddings did not prove insightful in knowledge discovery about the violence severity and its impact on the victims.

From Table 5.8, the most common potential sources of misclassifications have been identified and classified as (i) misspellings (e.g. 'violenc', 'voilence'), (ii) abbreviations (e.g. 'dv', 'dvam', 'gbv'), (iii) synonyms (e.g. 'violence', 'assault') and (iv) phrases (e.g. 'stop the silence', 'love dont hate', 'heal dont hurt'). The embeddings used for model training effectively address the misclassifications concerns by accounting for the semantic relationships between the terms, as represented by their cosine similarity.

As DV embeddings are trained on the posts collected from platforms where victims share their stories and seek support, the insights obtained prove invaluable for public health monitoring and suitable preventative measures design.

5.5 Summary of Findings

Social Media has been increasingly used in violence prevention by awareness raising, knowledge sharing, and bringing stories to the public [43]. Despite the

increasing popularity of self-disclosure and support seeking among DV victims, the limited research exists with regard to the actionable insights extraction in DV domain. Given the large volume and unstructured format of Social Media data, the robust and scalable posts classification techniques development proves essential in the efficient content management and timenly intervention by DVCS groups moderators.

Thus, the approach for *multi*-class identification from DV Social Media posts with the state-of-the-art Deep Learning models for the support of DVCS groups has been proposed. The main contributions are as follow:

(1) Medium-scale benchmark DV dataset with multi-class annotation construction 'gold standard'; (2) Deep Learning classification model development and its performance evaluation against its various architectures; (3) Deep Learning model performance validation against the selected Machine Learning baselines; (4) Visually-enhanced interpretation of the similarities between the categories and the main sources of misclassifications; (5) Domain-specific embeddings construction and its evaluation from the classification improvement and insights generation point of view.

An importance of the availability of annotated corpora to reduce the time and costs involved in manual human annotation process in the future is emphasised in [212]. It is particularly relevant to the niche applications such as DV. Given no previous work on the required fine-grained level of detail in the context of DV, the 'gold standard' dataset under the supervision of the domain expert has been created.

A comprehensive set of experiments, covering all possible 'feature-model'

combinations has been performed with the results specified in Table 5.4. On average, the Deep Learning models with words embeddings obtained higher performance in comparison with the traditional Machine Learning approaches (except for RNNs). The maximum scores were obtained for GRUs with GloVe words embeddings, Nadam optimizer and a batch size of 32. Thus, the application and optimisation of various Deep Learning architectures as the practical solution to real-world problem was demonstrated along with the empirical validation of its superiority over the traditional Machine Learning techniques.

As Deep Learning is highly advanced computational technique, the issues may arise with regards to the subsequent results interpretation. The dimensionality reduction scatter plots provided an intuition behind each model classification performance through categories separation in 2D space. The confusion matrices further complemented the analysis by quantifying the classification scores for each group as well highlighted the main sources of misclassifications. As a result, the BLSTMs proved advantageous in the case of *Personal Story* (92% 5.4b), whereas GRUs in the case of *Fund Raising* and *Awareness* (97% and 89%, respectively 5.4c). The decision-support regarding the most optimal model selection for the particular class distinction was therefore provided.

The advantage of domain-specific words embeddings has already been proved in literature, e.g. crisis embeddings [150] and medical embeddings [251]. Given the expected classification performance improvement and the potential for Knowledge Discovery, the DV-specific embeddings have been generated. The classification accuracy of Deep Learning models was marginally higher with DV embeddings and the low number of dimensions (50), which proves beneficial considering the reduced training time. Finally, the words analysis

from the domain-specific embeddings enabled to obtain fine-grained insight into the abuse types as well as the health conditions experienced by the victims. In contrast, the results from GloVe proved generic and little informative from DV perspective.

Still, the findings presented should be considered in light of the several limitations. The size of the corpus was considered moderate (1,655 posts), due to laborious manual annotation process. Nonetheless, the posts distribution among the classes was relatively similar (*Awareness*-20.9%, *Empathy*-22.4%, *Fund Raising*-17.4%, *Personal Story*-21.3%, *General*-18.0%, and proved sufficient for model training and categories identification. Additionally, the words embeddings approach naturally extends the features vectors, effectively leveraging even small scale datasets. The benefit of posts collection from platforms other than Facebook was recognised as well. As a result, the analysis with respect to particular source of data would further enrich the study (e.g. What is the classes composition across the platforms?). Also, the on-going monitoring of DV-related Social Media discourse would enable an identification of the emerging new categories. Despite the limitations identified, the step towards pro-active support and mitigation of the destructive impact of DV on physical and mental health of its victims with state-of-the-art technology has been proposed.

CHAPTER 6

AUTOMATIC IDENTIFICATION OF ABUSE TYPES AND HEALTH ISSUES FROM ONLINE DOMESTIC VIOLENCE POSTS USING DEEP LEARNING

DV is emerged as a growing concern due to the serious threat, it poses towards public health and human rights. Victims suffer from various health issues such as emotional, physical and reproductive, that are acute, long-lasting and chronic, or fatal. In recent years, the availability of Social Media has allowed DV victims to share their stories and to receive support from community. Though, the sheer volume of user-generated data has become useful resource for public health monitoring, it has few challenges as the user posts are informal, the medical terms are nontechnical and difficult to extract from lengthy posts. The aims of this chapter are: *(i) to present three datasets that we prepared for the task of identifying abuse and health mentions from user posted Social Media data; (ii) to automatically predict the various forms of abuse and health mentions, from the victims' posts using advanced Deep Learning technique; (iii) to investigate that, the multi-corpus training approach improves the classification performance and (iv) to investigate that, the trained word embeddings using a large DV corpus capture the precise semantic and syntactic word relationships than the existing large corpora of pre-trained word embeddings.* Empirical evidence on a ground truth datasets has achieved an prediction performance of up to 78% on average. This chapter validates an application of cutting edge technique to a real-world problem and proves beneficial to DVCS groups, health care practitioners and most of all - DV victims.

6.1 Introduction

As discussed in Chapter 2, DV involves violent, abusive or intimidating behavior by a partner or a family member to control, dominate or cause fear to other family member(s) [162]. According to WHO estimates, 35% of women worldwide have experienced violence from an intimate male partner [76, 164]. What is more, IPV spans a wide range of abuse types i.e. physical, sexual, psychological, verbal and even financial. IPV has got serious and durable impact not only on victims's and their children's health and well-being, but also on a society as a whole.

According to VicHealth [238], IPV contributes to more death, disability and illness incidents among women aged 15 to 44 than any other preventable risk factor, as per Burden of Disease methodology. Using the similar methodology, the research survey findings state further that IPV constitutes a significant risk to women's health leading to physical injuries, mental illnesses and reproductive disorders [240]. According to the statistical results of International Violence against Women Survey, two in every five women experienced physical injuries in the incident of violence [146]. Physical injury is common as a result of IPV. The most common injuries were bruises, swelling, cuts, scratches, burns, broken bones or head and brain injuries. IPV also have a cumulative impact on a women's mental health [195]. Victims reported increased rates of severity and co-morbidity of mental disorders, and higher level of mental-health related dysfunction. Based on the Australian Longitudinal study on Women's Health [9], IPV causes poorer mental health issues such as higher rates of PTSD, depression, anxiety, suicidal ideation and inflicted self harm. Not only physical and mental health, it has serious implications on reproductive health and more likely to

experience miscarriage and sexually transmitted diseases.

The high costs associated with IPV highlight an urgent need for the effective and issue-specific initiatives development and, as a result, proactive abuse instances prevention. The goal can be achieved only through the sound and evidence-based evaluation of the health concerns reported by the victims. The action towards women and children at risk situation's improvement starts from the awareness. The more knowledge is available to the decision-makers, the greater the chance of the successful and timely violent acts' prevention.

The number of violence prevention and mitigation initiatives is currently increasing, given the scale of the problem, and the severity of its consequences. According to Ellis et al. [61], disclosure of emotional based traumatic exoeriences can help to connfront mental illness. Social Media platforms offer a robust source of DV posts, as victims increasingly share their personal abusive experience on the online DVCS groups. It has been found that the emotional support received from formal (e.g. DVCS groups) and informal (e.g. family, friends) sources, commonly referred to as 'having someone to talk to', has positive impact on individuals mental well-being [227] [204]. For instance, consider the post shared by DV victim, *'I desperately need help. He physically assaulted me and threatened to kill me. I have spent the last 10 months with depression, and PTSD'*. The details of violent incidents *(physical, emotional etc.)* as well as the related health issues before/after exposure to abuse *(depression, trauma etc.)* are shared by the victims. Thus, the timely detection of potential DV incidents are indispensably important for DVCS groups to pro-actively reach out the victims in a timely manner. However, the abuse and health mentions are embedded in lengthy descriptive user posts and an effective technique is required for automatic mining

of Abuse Types and Health Issues (ATHI). But the victims from diverse cultural background and linguistic preferences generates the posts with diverse expressions, misspellings, abbreviations, and ambiguous mentions. Consider the following examples. (i) The term 'Domestic Violence' is represented in varied forms such as 'DV' or 'Domestic Abuse'. (ii) The term 'physical' abuse is misspelled as 'pysical' or 'pyhsical' abuse. (iii) The health issues are often mentioned in abbreviations such as 'ocd' (Obsessive Compulsive Disorder), and 'ptsd' (Post-Traumatic Stress Disorder). These varied expressions called noisy data pose serious processing challenges, in producing meaningful results.

Considering the limitations listed, the study proposed employs the state-of-the-art Deep Learning algorithms to automatically extract ATHI mentions from Social Media posts of DV victims. An alternative approach to survey-based design has been demonstrated that utilises the abuse reports posted in risk-free online environment. The advances in Deep Learning and natural language processing techniques provides great opportunities to mine the highly unstructured and voluminous data for analysis and actionable insight. The framework provides the scalable and standardised approach to proactively support violence victims as well as facilitates the Burden of Disease estimation due to increased data granularity.

In our previous chapter (Chapter 4), the approach for binary classification of 'critical' versus 'non-critical' online posts using Deep Learning has been proposed. The primary objectives of this chapter are to design, implement and validate Deep Learning approach to automate relevant information extraction from unstructured text-based Social Media data. We demonstrate that our approach can effectively capture a valuable information regarding the patterns of victims'

abusive incidents and associated health states in a highly efficient and systematic manner. To the best of our knowledge, no prior work has either focused on extracting fine-grained knowledge from online posts related to DV area nor evaluated Deep Learning techniques against various feature extraction methods. The experiment results are beneficial to the researchers, who are interested in conducting further research in violence prevention domain. Taken together, our approach opens new opportunities for scalable and automatic information extraction from unstructured DV posts. The main contribution of the study are as follows:

- DV corpora creation with ATHI annotation from multiple data sources;
- Application of Deep Learning models for informative feature extraction (ATHI);
- Performance comparison of state-of-the-art Deep Learning models;
- Multi-corpus training of distinct corpora to investigate the performance improvement;
- Domain-specifiv embeddings versus default embedding (GloVe) performance analysis;
- Knowledge discovery about the violence issue from Social Media.

The rest of this chapter is organized as follows. Section 6.2 provides the background on health-related knowledge extraction from Social Media. Section 6.3 explains data extraction and gold-standard label annotation. Section 6.4 describes about feature extraction and model construction for automatic identification of ATHI. Section 6.5 provides details on experiments to evaluate our approach the proposed framework, with analysis of the results and discussion. Section 6.6 concludes the chapter and envisages future research directions.

6.2 Mining Web for Health-related Knowledge Extraction

Due to its successful past applications, Social Media and other online sorces have become a popular outlet for health-related data mining. As the traditional data collection approaches tend to under-report the actual public health concerns, the demand for alternative data sources emerged. Given the expressive nature of Social Media and its propensity towards human interactions fostering, its use for knowledge mining about the societal issues from directly affected users was only the matter of time. Addtionally, the numerous medical forums offering health support have come into existence and their popularity only keeps growing.

One of the most common examples in literature considers mining online posts for Adverse Drug Reactions (ADRs) and it has been studied since 2010. Leaman et al. mined patients' comments on medical forum *DailyStrength* [54] to find mentions of ADRs. The data was annotated for adverse effect, beneficial effect, indication and other. The ADR lexicon has been utilised. Furthermore, the study on pharmacovigilance from Social Media [156] has observed that 'while a few individuals' experiences may not be clinically useful, thousands of drug-related posts can potentially reveal serious and unknown ADRs. As a result, the extended data sources such as Social Media or health-related forums 'augment the current systems' [156] and effectively the extraction of complex medical concepts with relatively high performance from informal, user-generated content has been validated [34, 38, 156, 212].

As traditional Machine Learning models require human experts to encode domain knowledge through feature engineering, Deep Learning approaches

are able to learn salient feature representations and achieve state-of-the-art results [75]. This characteristic makes them especially apt for Natural Language Processing tasks, where manual encoding of features is both impractical and inefficient [75, 118]. Various applications of Deep learning in Natural Language Processing context range from sentence modelling [108] through text classification [113] to topic categorisation [105].

There have been numerous studies that successfully utilised Deep Learning for health-related knowledge extraction. In [75], the clinical pathology reports have been leveraged in the context of cancer surveillance to abstract essential information regarding tumor type, location and histological grade. They proved that a well-designed automated solution could standardise how tumor data is encoded (human bias addressed) and yield improvements in the accuracy overtime given the growing dataset size (scalability aspect addressed, which is often the case when dealing with online generated content).

By far, no attempt has been made to design, develop, implement and validate Deep Learning techniques for knowledge extraction in DV domain. Specifically, the ATHI (that is of particular value for Burden of Disease estimation) automatic identification from highly unstructured textual data is still unexplored in literature.

Table 6.1: Descriptive Statistics of 3 Data Sources

Statistics	Dataset	Facebook	Reddit	Blogs
Posts information	No of collected posts	200k	25k	50
	No of annotated posts	225	230	50
	Metadata collected	*post_comment, created, id, score, title, upvotes, url, timestamp, comments_number*	*post_comment, created, id, score, title, upvotes, url, timestamp, comments_number*	*blog_link*
No pre-processing	Total No of words	92668	88337	55886
	Max words count of posts	4387	2390	6031
	Avg words count of posts	413	383	1124
	Most common words	I, to, and, the, my, a, was, of, he, me, in, that, for, it, with, is, have, had, this, on	I, and, to, he, the, my, me, a, was, that, of, it, him, for,in, but, with, so, is, have.	to, I, and, a, my, the, of, by, support, here, click, in, is, her, for, was, domestic, she, help, with.
Stop words removal	Total No of words	42376	37543	24351
	Max words count of posts	2012	985	2722
	Avg words count of posts	188	163	487
	Most common words	Would, one, get, time, never, like, life, abuse, years, know, back, still, got, violence, could, children, go, help, family, even.	Would, like, get, know, time, feel, told, even, want, things, one, relationship, never, back, still, said, go, abusive, got, going.	Would, time, like, get, one, know, back, life, going, got, could, go, things, never, abuse, even, home, friends, day, family

6.3 Benchmark Dataset Construction

6.3.1 Data Extraction

The data has been extracted from three popular Social Media platforms, namely *Facebook*, *Reddit*, and *DV-related blogs*.

- Our first dataset has been sourced from *Facebook*, the popular Social Media site. Facebook gathers approximately 2 billion users worldwide and is ranked first among the top 15 social networking sites [191]. The posts, comments, and associated metadata from Facebook groups, that discuss the range of DV issues, through Facebook Graph API [1] with the search term of 'Domestic Violence and Domestic Abuse'. A number of posts and comments of approximately 200*k* was returned following the data collection from the 10 most active DV pages from the time span of 2011 to 2018.
- To further extend the scope of the analysis, another highly popular social

[1] https://developers.facebook.com/docs/graph-api

support platform - *Reddit* was considered. Reddit attracts as many as 542 million monthly visitors [248]. According to Choudry et al., [56], the distinctive feature of Reddit such as account creation within minutes without the provision of an email address (so called 'throwaway' accounts), makes it a particularly popular outlet for sensitive information disclosure. Reddit also allows for formation of sub-communities called "*Sub-Reddit*" that concentrate on more specific issues. Given the prevalence and significance of an impact the abuse instances have got on a society, the numerous Sub-Reddits have been formed to provide support for victims, survivors, their family and friends. Similarly as in the case of Facebook, the posts, comments and associated metadata have been collected. The following Reddit API has been used [2]. The corpus statistics are detailed in the Table 2. Nearly, 25k posts and comments was retrieved from 15 active sub-Reddits with the time span of 2014 to 2018.

- The third and final dataset has been sourced from the social platform - *blogs*. Blogs are considered as an important platform for sharing informational and emotional thoughts and freedom of individual expression [106]. Some of the top nonprofit organizations such as 'Domestic Violence Resource Centre Victoria' and 'Steps to End Domestic Violence' not only work to prevent and respond to DV, but also share stories from women, who have experienced abusive relationships for awareness promotions. Hence, we collected nearly 50 personal stories from various blogs.

Considering the ethical concern, only the public posts are considered from Facebook groups and Sub-Reddits. Furthermore, the identity of individuals from the dataset is not disclosed and the content is mined solely for the aggregate knowl-

[2]https://www.reddit.com/dev/api

edge extraction purpose. The insights about the corpus characteristics were produced such as number of collected and annotated posts with metadata information, the maximum and average words count, before and after pre-processing. Also, the most frequent words were produced for qualitative analysis. Table 6.1 shows that, the number of annotated posts is very few, despite the higher volume of posts collected. This is due to the fact of larger vocabulary size of posts. The vocabulary size of posts is nearly 55886 for just 50 collected posts of Blogs. Similar in the case of Facebook and Reddit, the vocabulary size is much larger i.e., 92668 and 88227 respectively.

When the most common words were analyzed and most of the 20 frequent words found to be personal pronouns, prepositions such as *'I, me, my, he, she, him, and her'*. These words are called as stop-words and considered to be insignificant, in general. However, in our DV context, the stop-words add value to the knowledge discovery and was proven significant in our previous findings of DV critical posts identification [223]. For e.g. Consider the post *'I was in abusive relationship for 10 years. He physically assaulted me'*. In the given example, the proportion of stop-words are significantly higher than the general words and deemed necessary, considering the self-expressive and descriptive nature of DV posts.

The stop-words contribute to nearly 40% of the total words count. Further analysis was made after stop-words removal, to gain some insights about the specifics of other DV related words. The words such as *'abuse, violence, abusive, relationship'* were more prevelant and related with the violence situation of the victim. The word *'children'* was widely used, as they act as the primary victims in many DV cases. The terms *'friends and family'* are usually used to represent

the supporters in critical situations. The terms relating to the time/length of the abusive incidence such as *'years or time'* have also been observed. As presented, the analysis of the total count of words before and after pre-processing and the most frequent words has found interesting in terms of the potential insights and knowledge extraction.

6.3.2 Gold Standard Labels Construction

Following the data collection phase, the next stage is to construct a benchmark dataset [3] for evaluating the proposed approach. The posts shared on online DV groups do not only contain victims' personal stories that could be mined for valuable insight, but also cover awareness promotion campaigns, advertisements, empathy expressions etc. Further details are described in our previous work on 'automatic classification of DV crisis posts' [223]. As *Personal Story* category has been proven critical as a result of help-seeking behaviour emerged in the content. Hence, we used the 225 *Personal Story* posts from Facebook, 230 posts from Reddit and 50 posts for the ATHI mining. Despite relatively small sample size, the number of tokens extracted amounted to as many as $2,36,891$ due to the lenghty and highly descriptive nature of DV-related *Personal Story* posts.

In terms of data annotation, the two human scorers under the supervision of a consultant psychiatrist specialized in DV discipline as well as anxiety and depressive disorders. Each post (all Facebook, Reddit, and blogs dataset) was labelled at the individual token level for the presence of Abuse Type and Health Impact. To validate the inter-rater reliability of the scorers, the Kappa coefficient

[3] https://github.com/sudhasmani/DV_Dataset

Table 6.2: Distribution of Annotated Instances in Each Corpus

Corpus	Facebook	Reddit	Blogs
No of Health Instances	807	784	385
No of Abuse Instances	2539	1881	1197
No of Non-ATHI Instances	89322	85672	54304
Total	92668	88337	55886

[136] was calculated. The achieved degree of agreement was reasonably high at 0.81. Any disagreements were resolved by the domain expert in order to ensure the correctness of the classification process.

For example, the posts P_{1-3} in Table 6.3 were shared by victims and disclosed the abusive incidents in their relationship. It can be noticed that the victims' describe various particular Abuse Types caused by perpetrators. These include: *(1) 'Physical Abuse' (beating, hitting, shoving); (2) 'Verbal Abuse' (screaming, yelling); (3) 'Emotional Abuse' (threatening); (4) 'Sexual Abuse' (raping) and (5) 'Financial Abuse' (devastated financially). In addition, the victims' posts further reveal the various Health Issues that can be (1) 'Physical' (bruises, bleed); (2) 'Mental' (Depression, Anxiety, Trauma, PTSD) or (3) 'Reproductive' (pregnancy, abortion).* Table 6.4 further lists the most frequent words associated with the ATHI mentions.

Table 6.2 provides the summary of annotated instances. Note that, the proportion of annotated ATHI instances was much lesser than non-ATHI instances in all the 3 annotated corpora. For e.g. the ATHI instances comprised of just 3.6%, 3%, and 2.8% for Facebook, Reddit, and Blogs respectively. Whereas, the proportion of non-ATHI instances were much higher to 96% on average in all the corpora. Within the ATHI instances, the health issues were identified to be minimal in number than the abuse types. The reason being, health issues were not mentioned in the victims' descriptive posts as often as the abusive experience. This kind of higher difference between the labeled entities was even

observed in the literature [251] for ADR extraction (Drug & ADE entities were just 3% & 5%, whereas the other entities were nearly 92%) from Social Media posts. The reason observed, as symptoms or reason for taking the drugs were not mentioned often in the patients posts [110].

Table 6.3: Examples of DV posts and the Corresponding ATHI Mentions

ID	DV Posts	Abuse Types	Health Issues
P_1	*"Within two weeks of my marriage, my husband was beating and raping me. Over the next seven years, he covered me with bruises. I got out when he started hitting our children. I survived, I'm safe, and I can talk about it now. "*	*beating, raping, hitting*	*bruises*
P_2	*"It's kinda dark. In the beginning every relationship is almost perfect. But remember, the honeymoon stage doesn't last forever. Well. Mine sure didn't. He started yelling at me. It hurts. I was scared. I went to go give him a hug. He got up grabbed me slapped me in the face then punched me in the stomach. I was shocked and still suffering from depression and anxiety attacks. "*	*yelling, grabbed, slapped, punched*	*hurts, scared, shocked, depression, anxiety attacks*
P_3	*"I am a single mother surviving domestic abuse, stalking and harassment, leaving me while pregnant. I have been hit, shoved, and slapped. I had nightmares, I was nervous all the time. I found a therapist that dealt with my trauma/ptsd. "*	*domestic abuse, stalking, harassment, hit, shoved, slapped*	*pregnant, nightmares, nervous, trauma, ptsd*

6.4 Model Construction

This stage presents the proposed approach for automatic identification of ATHI mentions from DV posts. We adopted two popular state of the art Deep Learning architectures for ATHI extraction task. The Deep Learning algorithms rely on the feature extraction part, that uses word representations as the features.

Table 6.4: Most Frequent Words associated with ATHI Mentions

ATHI mentions	Most frequent words
Abuse Types	abuse, violence, domestic, physical, hit, broke, threats, sex, kill, emotional, fight, beat, control, throw, push, blame, grab, mental, scream, yell, kick, punch, cut, verbal, manipulate, pull, financial, slap, rape, argument, assault, attack, stalk, choke, accuse, destroy, drag, death, harass, suicide, strangle, isolate, smash, trap, trouble.
Health Issues	hurt, scared, fear, pain, pregnancy, bruise, depression, afraid, anxiety, sick, suffer, stress, sad, trauma, blood/bleed, nightmare, mental illness, panic, worry, shock, ptsd, miscarriage

Two different kinds of feature representation were used in this work such as word and character level embeddings, when training the Deep Learning model. They are discussed in detail in the following sub-sections.

6.4.1 Feature Extraction

As the language in Social Media is highly informal and user-expressed concepts are often non-technical, descriptive and challenging to extract [156], the word embeddings have been used to account for the limitations identified. The word embeddings are real-valued vectors that capture words' similarity, and eventually improves the prediction performance on unseen or rarely mentioned instances. The vectors are learnt in such a way that words with a similar meaning will be positioned nearby in the vector space. Thus, the embedding approach overcomes non-semantic representation shortcoming as well as data sparsity present in the traditional textual feature representation techniques (called Bag of Words or One-Hot encoding).

The two most common word embeddings that were trained on the large external corpus such as Google's Word2Vec [142] and Twitter's crawl of GloVe

[181] have already shown promising results in various tasks such as mining adverse drug reaction mentions from Social Media posts [228, 259]. However, these classic word embeddings work well, when they are applied in a large text corpus. When the corpus size is smaller or sparse in any specialized domains (e.g. cybersecurity [205], disease surveillance [79]), they failed to generate high quality vectors. Considering this reason, domain-specific word embeddings were trained on available domain related corpus, which are usually sparse. The domain-specific embeddings were also applied and validated, demonstrating the improved performance in text classification (crisis embeddings [150]) and adverse drug reaction identification in health Social Media posts (medical embeddings [251]). Domain-specific embeddings not only proved to leverage domain vocabulary, but also improve semantic relations in domain related texts of smaller corpus size [205].

In order to investigate the impact of domain-specific embeddings in ATHI prediction performance improvement, DV embeddings have been constructed firstly. Secondly, the prediction accuracy of Deep Learning models trained on pre-trained and DV-specific embeddings was then compared. The details of embeddings and experiments performed are as follows:

Pre-trained embeddings: The two most popular embeddings namely Word2Vec and GloVe have been used in our previous work for DV critical posts identification [223]. As the Deep Learning models appeared to achieve better performance with GloVe than with Word2Vec in our previous work, we preferred to experiment with the GloVe embedding for ATHI prediction. Both 300 and 50 dimensions were included for comparisons, and each word is represented by a vector of word embeddings containing 300 and 50 dimensions. The

former has been trained on nearly 840 billion tokens from Twitter posts, and covers 300 dimensional vectors for a vocabulary set of 2.2 million words and phrases [181]. The latter one is trained on nearly 6 billion tokens and covers 50 dimensional vectors for 400k words.

Domain-specific embeddings: The domain-specific embeddings have been trained on the large corpus of DV-related discussions to differentiate from the generic news and tweets. The sources for data extraction included Reddit, Twitter, Facebook, and blogs and only topic-relevant posts were considered (e.g. victims support forums, abuse-dedicated groups etc.). In total, the corpus contained nearly 500k posts. The 50 embedding dimensions were used for training, given the relatively smaller corpus size in comparison with the pre-trained embeddings. As the domain-specific knowledge are difficult to collect and define, we evaluated the usefulness of domain-specific embeddings induced from available textual resources in the DV context.

Character embeddings: Word embeddings is state of the art on dealing with Natural Language Processing tasks. However, there is a potential shortcoming associated with Out-of-Vocabulary (OOV) issue. OOV issue more commonly occurs with the clinical posts or domain-specific datasets, when the terms not appear in the training data, but only appears in the test data. Word embeddings can't handle those unknown/rare words. For e.g. consider the abuse related terms {*'physical, emotional, sexual, financial'*} or {*'beating, battering'*} in the DV context. Similarly, both the abbreviated terms {*'ptsd, cptsd'*} related with health issue. OOV issue might occurs, if the model hasn't seen a term during training. As those terms shared the common suffix and exploiting those prefix and suffix information about the words can help to tackle the OOV issue. Character

embeddings [261] are used to encode character level features of the words that share similar prefix and suffix and obtain closer representations among words on the same category based on their characters. The character embeddings neither require knowledge about the words nor their syntax/semantic structures. In general, the explicit character level features such as prefix and suffix are effectively utilized [210] to tackle the OOV issue. Finally, the comprehensive word representation was used in our approach, which combines word embedding features with character level information.

6.4.2 Model Specification

Two Deep Learning models were constructed and compared at this stage to extract ATHI mentions from online DV posts, namely:

BLSTMs: We adopted BLSTMs model [86] for ATHI identification task, as it proved to be powerful and flexible type for NER tasks, that deal with sequence classification. LSTM cells in general, process the information in one direction as feed forward network. But, the BLSTMs process the information in both forward and backward directions and learn the whole sentence by exploiting the information of both short and long-term dependencies of a word. Thus, the BLSTMs layer forms the core of the network and has the following three inputs such as character level patterns, word level input (either from pre-trained ones or domain-specific embeddings) and PoS tagging features. The embed layer of BLSTMs creates both character and word embedding, by mapping the characters and words with real numbers.

Stacked Residual BLSTMs: The stacked residual BLSTMs architecture used

is described in-detail in [230]. Layer stacking is a traditional way of adding more representational power to a neural network model, where the upper layers can learn to compensate the representation errors of information from the lower-level inputs. However, stacking multiple layers suffers from degradation problem due to the difficulty in training multiple layers and fit these layers to the underlying mappings in a desired format [89]. To overcome this limitation, the residual connections between stacked layers is introduced in stacked residual BLSTMs model [230], where the upper layer of a neural network has direct access to the original input from the output of lower layers without any dimensionality reduction. This model achieved the state-of-the-art results for NER task of both datasets on Spanish and English languages.

6.5 Experimental Design and Results Analysis

In this section, the experiments for automatic identification of ATHI instances from online DV posts are discussed in detail. Several steps were performed to evaluate the performance of the introduced approach using Deep Learning. These include:

(A) *Model Training:* The detailed steps for model training are presented including features extraction approaches as well as the settings selection, and the rationale behind their application.

(B) *Accuracy Evaluation:* The performance of the 2 Deep Learning models, namely BLSTMs and stacked residual BLSTMs on the constructed benchmark data set was evaluated. The most commonly used validation metrics, i.e. Precision, Recall, and F-Measure were calculated.

(C) *Multi-corpus Training Results:* The training dataset from distinct corpora were combined due to the largely imbalanced dataset and to investigate the improvements in classification performance.

(D) *ATHI Recognition and Error Analysis:* The sample posts from the test dataset were identified and the correctly classified ATHI mentions were presented to demonstrate the efficacy of Deep Learning technique. In addition, the possible reasons behind classification errors were also discussed with some example posts, following the possible solution to address the issues identified.

(E) *Domain-Specific Embeddings Analysis:* The experiments were conducted to test our hypothesis of DV-specific embeddings over the generic pre-trained embeddings performance improvement. The analysis covered (i) an impact of the proposed embeddings on prediction performance, and (ii) the insights and knowledge discovery about DV from the embeddings generated.

6.5.1 Model Training

We experimented the two Deep Learning models with the constructed benchmark datasets for ATHI prediction task. Since, three datasets were involved, the train-test split approach was used, as widely applied in previous works for NER recognition task [156, 212, 251]. Thus, we divide each of the three datasets into two parts: 90% for training and 10% for evaluation. In order to examine the robustness of the classifiers, the features for Deep Learning models were extracted using the GloVe word embeddings of 300 and 50 dimensions. The first layer of the model is the embedding layer that computes the index mapping for

all the words in the vocabulary and convert them into dense vectors of fixed size by parsing the pre-trained embedding. In addition to the word embedding features, the character level features and PoS tagging were extracted and fed into the BLSTMs layer of 100 units. Additionally, the models were trained up to 50 epochs and implemented using Keras [4]. Number of training epochs were set to 50 and Nadam optimizer was used, as popularly used in previous applications [196]. To avoid overfitting, the dropout was added on hidden units in all layers and set to 0.6, following the higher performance in existing works [35]. SoftMax activation function was implemented at the output layer, as classifier to predict the ATHI mentions. W.r.t Stacked residual BLSTMs model, the layers were added in addition with the aforementioned settings. Though the significant performance improvement was improved with more stacked layers [241] [230], we limited the experiments with 2 stacked layers considering the computational complexity and medium scale benchmark dataset. The residual connection is inserted between the two stacked layers of BLSTMs model and dropout is added after residual connection to make use of its regularization effect.

6.5.2 Accuracy Evaluation

Various evaluation metrics such as Precision, Recall, and F-Measure were computed for both Deep Learning models and are presented in Table 6.5. Overall, the Deep Learning models with GloVe embedding of 300 dimensions achieved improved performance over 50 dimensions for all the three datasets. From Table 6.5, it is obvious that the stacked residual BLSTMs achieved higher per-

[4]https://github.com/fchollet/keras

Table 6.5: Performance Evaluation of Deep Learning Models with Pre-trained Embeddings

Model	Embedding	Corpora	Instance Type	Precision	Recall	F-Score
Stacked Residual BLSTMs	GloVe (300 Dims)	Reddit	Abuse	0.89	0.78	**0.83**
			Health	0.93	0.58	**0.75**
			Average	0.91	0.68	**0.79**
		Facebook	Abuse	0.88	0.74	**0.81**
			Health	0.95	0.54	**0.74**
			Average	0.91	0.64	**0.77**
		Blogs	Abuse	0.88	0.76	**0.82**
			Health	0.94	0.55	**0.74**
			Average	0.91	0.65	**0.78**
	GloVe (50 Dims)	Reddit	Abuse	0.73	0.59	0.66
			Health	0.85	0.48	0.66
			Average	0.79	0.53	0.66
		Facebook	Abuse	0.68	0.53	0.60
			Health	0.78	0.45	0.61
			Average	0.73	0.49	0.61
		Blogs	Abuse	0.70	0.55	0.62
			Health	0.82	0.47	0.64
			Average	0.76	0.51	0.63
BLSTMs	GloVe (300 Dims)	Reddit	Abuse	0.86	0.62	0.74
			Health	0.90	0.51	0.70
			Average	0.88	0.57	0.72
		Facebook	Abuse	0.84	0.64	0.74
			Health	0.92	0.51	0.71
			Average	0.88	0.57	0.72
		Blogs	Abuse	0.85	0.62	0.73
			Health	0.89	0.50	0.69
			Average	0.87	0.56	0.71
	GloVe (50 Dims)	Reddit	Abuse	0.72	0.57	0.64
			Health	0.82	0.42	0.62
			Average	0.77	0.49	0.63
		Facebook	Abuse	0.64	0.46	0.55
			Health	0.73	0.40	0.56
			Average	0.68	0.43	0.55
		Blogs	Abuse	0.68	0.55	0.61
			Health	0.72	0.42	0.57
			Average	0.61	0.48	0.59

formance than BLSTMs model. For instance, w.r.t 300 dimensions of GLoVe, stacked residual BLSTMs model achieved nearly 78% accuracy on average for all 3 benchmark datasets, whereas the BLSTMs model achieved only 72% on average. Thus, the stacked residual connections significantly improved the performance across all three datasets. Overall, stacked model avoids degradation issues, with no increase in the computational complexity with 2 constructed layers. However, while the models achieved higher F-Score value on average, the recall value is comparatively lower for ATHI mentions and in particularly with health issues. One obvious reason for this, could be the lower number of annotated instances for ATHI mentions. Thus, more annotation is required to improve the size of the annotated ATHI instances in the dataset, which is expensive process. Hence, we further investigated to address the issue of dataset size, by performing multi-corpus training (sub-section 6.5.3), as suggested in the previous work [212].

6.5.3 Multi-corpus Training and Prediction Results

In the ADR extraction task, it is common to have largely imbalanced dataset, where there was lower availability of ADR instances available in contrast to the higher availability of non-ADR instances. Larger volumes of data might require increasing the number of ADR class instances and to train the algorithm. However, this is considered as an expensive process for preparation and annotation of data. Hence, Sarkar et al. [212] introduced an approach of multi-corpus training to improve the classification performance, where the dataset from multiple sources were combined for training, to classify the test instances for each of the source. They performed two different types of experiments, (i) only the ADR

Table 6.6: Paired Performance Evaluation for ATHI Mentions over the Three Corpora

Test data	Training data	precision	recall	Fscore
Reddit	Reddit + Facebook	0.93	0.73	0.83
	Reddit + Blogs	0.91	0.70	0.80
Facebook	Facebook + Reddit	0.93	0.71	0.82
	Facebook + Blogs	0.92	0.70	0.81
Blogs	Blogs + Reddit	0.90	0.71	0.80
	Blogs + Facebook	0.90	0.69	0.79

instances from another dataset were added to training dataset, to increase the number of minority classes i.e., ADR instances. (ii) all the instances from another dataset were added to training dataset, to increase the overall volume of the dataset. Hence, we adopted the second approach by combining all the training instances i.e., (both ATHI and non-ATHI) from distinct corpora to investigate the classification performance. Because the first approach of adding only the ATHI instances from another corpora is not possible, as the Deep Learning algorithm analyses the entire post for possible ATHI identification task.

Stacked residual BLSTMs with GLoVe embedding of 300 dimensions performed significantly better than BLSTMs model. Hence, stacked residual BLSTMs were further used for multi-corpus training. The results of paired performance evaluation over three corpora for the average of ATHI mentions were shown in Table 6.6. The results shown that, the classification performances over the ATHI mentions significantly benefit from multi-corpus training considering the improved performance. Another advantage of multi-corpus training is that, all the paired corpora were from similar source i.e., Social Media and the results had complimented each other.

6.5.4 ATHI Recognition and Error Analysis

To further obtain the deeper insights into how the Deep Learning technique benefits the classification performance with respect to ATHI mentions, some of the posts from test dataset were sampled. The labels predicted by the Deep Learning model were compared with gold standard labels. As the maximum prediction accuracy was achieved by stacked residual BLSTMs model with pre-trained GloVe embeddings of 300 dimensions (78%), the exemplary posts with the correctly identified instances are presented in Table 6.7. The reference to either abuse types or health mentions is highlighted within each post. Even the posts were descriptive in nature, the Deep Learning technique accurately predicted the ATHI mentions. Based on the observation from the posts P_1 to P_3 of Table , the Deep Learning model is effective in identifying ATHI mentions, even with the presence of (i) abbreviations, (ii) non-standard expression (such as defined within quotes or brackets), or (iii) newly appeared terms during testing. However, the ATHI mentions of the posts P_4 and P_5 were not captured, due to the non standard terminology or spelling mistakes. Those misspellings, non standard or domain-specific terminologies were well captured by Deep Learning technique with domain-specific embeddings, which are discussed in detail in the next sub-section.

6.5.5 Domain-specific Embeddings Analysis

The domain-specific embeddings for DV were constructed and evaluated in terms of (i) impact on the prediction performance in Deep Learning models, (ii) useful insight generation and knowledge discovery about the DV. The com-

Table 6.7: Examples DV Posts with the Predicted Results and Error Analysis

ID	Posts	Correct Prediction?	Implication
P_1	*"My boyfriend is severely **emotionally abusive [AT]**. He's straight up told me that he **hurts [HI]** my feelings on purpose because "I deserve it" and says I don't deserve to be happy. Anyway, this paradox got me thinking about how my **(emotionally) abusive [AT]** relationship let me not think about myself..."*	Yes	'(Emotionally) abusive' was correctly identified as Abuse Types, even it was mentioned inside the brackets during the second time.
P_2	*"I've been trying really hard to get better . My therapist told me I have **PTSD [HI]** which was caused by my long time abuser. I don't know what's wrong with me but whenever I see something that reminds me of when he beat the shit out of me, I end up not being able to stop thinking about it..."*	Yes	'PTSD' was used in abbreviated form, however it was correctly identified as Health Impacts.
P_3	*"My marriage ended. I say marriage, I mean relationship. The divorce proceedings haven't started, because he **ummelled [AT]** me down to a shell of the character, I was before. We have 3 children in short span of time and he didn't help me out. no night feeds and no babysitting..."*	Yes	'Pummelled' term was not exist in the training dataset, but was correctly identified during testing as Abuse Types.
P_4	*"So I just left an **emotionally [AT]** and, in the end, **physically abusive [AT]** relationship. I feel lucky in that my family is extremely supportive and took me in immediately after things fell apart completely. However, I keep getting **flashbacks [non-HI]** of certain scenes I've experienced in this relationship and kind of relive the emotions I felt at the time..."*	No	'Flashbacks' was mis-classified as non-HI, considering the non standard terminology.
P_5	*"I have family and I hate my partner. We do have domestic issues and are currently in the process of going to therapy. I suffered from **physical voilence [non-AT]**... "*	No	'Phsyical voilence' was mis-classified as non-HI, as the term misspelled instead of 'physical violence'

Note: As the 'personal stories' posts were descriptive in nature, only the portion of the posts that define the ATHI mentions are specified here.

Table 6.8: Performance Evaluation of Stacked Residual BLSTMs with Domain-Specific Embeddings

Embedding	Corpora	Instance Type	Precision	Recall	Fscore
DV Embedding (50 Dims)	Reddit	Abuse	0.88	0.72	0.80
		Health	0.80	0.50	0.65
		Average	0.84	0.61	0.72
	Facebook	Abuse	0.91	0.72	0.81
		Health	0.77	0.52	0.64
		Average	0.84	0.64	0.73
	Blogs	Abuse	0.82	0.70	0.76
		Health	0.81	0.50	0.65
		Average	0.81	0.60	0.70

Table 6.9: Example Words and their Top 15 Similar Words from User Posts using DV and GloVe Embeddings

DV related words	DV embeddings	GloVe embeddings
Depression	anxiety, ocd, ptsd, insomnia, fibromyalgia, bipolar, schizophrenia, adhd, cptsd, disorder, bpd, severe, agoraphobia, psychosis, depressive	severe, illness, alcoholism, debilitating, anxiety, symptoms, experiencing, complications, suffering, caused, chronic, suffer, disease, infancy, ill
Anxiety	depression, ptsd, ocd, insomnia, adhd, flashbacks, disorder, anger, paranoia, psychosis, cptsd, fibromyalgia, panic, nightmares, bipolar.	persistent, experiencing, discomfort, stress, anger, pain, fatigue, paranoia, headache, symptoms, lingering, frustration, despair, disorder, confusion.
Physical	verbal, physically, psychological, emotional, physiological, pysical, severe, manipulation, scarring, aggression, phsyical, coercion, intimidation, beatings, battering	mental, psychological, experience, lack, stress, knowledge, certain, skill, learning, quality, reasons, aspects, effects, practical, behavior.
Sexual	exual, gendered, domestic, donestic, substance, dometic, sexualmisconduct, sexaul, interpersonal, believeme, ipv, preventable, intimate, spousal, unsilenced.	sex, abuse, behavior, harassment, sexually, rape, marital, discrimination, prejudice, parental, molestation, abusive, bullying, relationships, heterosexual.

Note: The words in the table also include misspellings and abbreviations, as they are more common in the users' postings on Social Media.

parative analysis with GloVe as a benchamark has been performed, and the moderating effect of different embedding dimensions was evaluated.

Table 6.8 presents the evaluation metrics of stacked residual BLSTMs model, trained on DV embeddings, with the dimensions set to 50. Only the stacked residual BLSTMs model was considered, due to its higher performance, described previously. In terms of the average prediction performance, the difference between GloVe and DV embeddings was negligible, which may not offset the time and cost involved in the domain-specific embeddings generation. The performance also varied across different dimensions levels, i.e. DV embeddings scored higher with the size of 50, whereas GloVe proved superior with the size of 300. From computational efficiency perspective, lower number of dimensions is preferable, given the reduced training time.

The other aspect of analysis involves the potential for insights generation

and knowledge discovery from both GloVe, and the domain-specific embeddings. DV embeddings were trained on the content crawled from online forums dedicated to the violence victims support. Hence, it is expected that the results obtained from domain-specific embeddings will be more meaningful and valuable, than the outputs from GloVe, trained on the general Twitter corpora. To test the assumption made, the 15 most similar words to the 4 DV-related terms were extracted from both GloVe and DV embeddings (Table 6.9). The similary was evaluated by the standard measure of cosine distance in the vector space. The 4 DV-related terms, in which 2 health mentions such as 'depression', and 'anxiety' and 2 abuse types such as 'physical', and 'sexual' were selected based on their common ocurrence in the user posts.

For instance, the sample words associated with 'depression' generated by DV embeddings included 'anxiety', 'insomnia', 'bipolar', 'schizophrenia' etc. whereas the words returned by GloVe included 'illness', 'symptoms', 'experincing', 'complications' etc. The domain-specific embeddings proved superior in the most prevalent post-abuse issues exploration. The associated health conditions detection also performed well regardless the commonly used abbreviations, such as 'ocd' (Obsessive Compulsive Disorder), 'ptsd' (Post-Traumatic Stress Disorder) or 'bpd' (Borderline Personality Disorder). On the other hand, the generic nature of terms produced by GloVe did not prove insightful in knowledge discovery about the health impact on the victims. Similar to 'depression', the sample words for 'anxiety' generated by DV embeddings also mined other relevant health related terms such as 'insomnia', 'flashbacks', 'paranoia', 'nightmares' etc.

As another example, the word 'physical' represents the type of abuse in the

domain context. The DV-embeddings return not only the instances of physical abuse (e.g. 'scarring', 'beatings', 'battering'), but also the other abuse types (e.g. 'verbal', 'psychological', 'emotional'). It also generates the mis-spelled words such as 'pysical', and 'phsyical'. In contrast, GloVe returns mostly irrelevant to DV terms such as 'experience', 'quality' or 'aspects'. Similarly to 'physical', the word 'sexual' and its related terms returned by the domain-specific embeddings allow for deeper insight into, in this case, the popular hashtags such as 'believeme' and 'unsilenced'. In addition, it also generates the abbreviations, such as 'ipv' (Intimate Partner Violence) and phrases (sexual misconduct).

The most interesting words returned by DV embeddings for the phrase 'sexualabuse' and the words are as follows: *'sexualassault, sexualviolence, childsexabuse, humantrafficking, childabuse, stopabuse, onechildistoomany, believeme, sexualharrassment, raisingawareness, childtrafficking, asylumseekers, metoochildren, kidstoo, csa, childbride, csasurvivors, gbv, rapeculture, breakingthesilence, ibelievesurvivors, healingandrecovery, victimblaming'*. Other than related phrases such as 'sexualassault', 'sexualviolence', 'childsexabuse', etc. it also generates some interesting and the most popular hashtags such as 'believeme', 'metoochildren', 'kidstoo' and the abbreviations such as 'csa' (Child Sexual Abuse) and 'gbv' (Gender Based Violence). But, in the case of GloVe embedding, an error message was produced as 'word sexualabuse not in vocabulary'.

From Table 6.9, the most common potential sources of mis-classifications have been identified and classified as (i) misspellings (e.g. 'pysical', 'phsyical', 'exual', 'sexaul'), (ii) abbreviations (e.g. 'ocd', 'ptsd', 'ipv'), (iii) phrases (e.g. 'sexualassault', 'sexualviolence', 'childsexabuse'), and (iv) hashtags (e.g.'believeme', 'metoochildren', 'kidstoo', 'breakingthesilence'). The em-

beddings used for model training effectively address the misclassifications concerns by accounting for the semantic relationships between the terms, as represented by their cosine similarity.

As DV embeddings are trained on the posts collected from platforms where victims share their stories and seek support, the insights obtained prove invaluable for public health monitoring and suitable preventative measures design.

6.6 Summary of Findings

The number of violence prevention and mitigation initiatives is currently increasingly, given the scale of the problem, and the severity of its consequences. The mark of abusive experience on the affected individuals, their families, and the society as a whole does not only result in traumatic psychological impact, but also poses enormous economic burden. The assistance and availability of the appropriate, well-trained, and specialised services is required in order to address the repercussions that DV has imposed. For the effective preventative strategies development as well as the tailored healthcare services provision, the accurate and real-time estimates are considered crucial. In particular, the most prevalent types of abuse (e.g. physical, sexual, emotional), the health conditions preceding the instance of violence (e.g. anxiety, depression), the health conditions suffered post exposure to violence (e.g. PTSD, insomnia, nightmares, flashbacks), or specific types of injuries experienced (e.g. bruises, bleeding). The popularity of detailed self-disclosures on Social Media for an informational/emotional/financial support throughout the traumatic experience has opened tremendous opportunities for valuable insights mining. Further-

more, the WHO has recognised the role of data on the circumstances surrounding the violent behaviours, as well as health-related information as crucial for pro-active victims support and associated economic costs estimation.

Motivated by the aforementioned, the framework for in-depth knowledge discovery about the ATHI has been proposed. Given the high-volume, high-velocity and high-variety of Social Media data, the robust and scalable ATHI prediction methods are considered vital not only for general awareness purposes, but also for the efficient content management and timely intervention by specialised support services, such as DVCS groups active on Facebook platform. Thus, the approach for ATHI identification from DV Social Media feeds with the state-of-the-art Deep Learning models for knowledge discovery as well as the support of DVCS groups has been proposed.

The three online sources (i.e. Facebook, Reddit and DV-related blogs), popular among the post-abuse victims were used for data collection. The descriptive statistics was performed on each of the corpora extracted for comparison purposes in terms of potential generation. The most frequent terms were pronouns and prepositions among all datasets, as expected. Despite being insignificant in general text mining applications, in DV context their inclusion was found important in 'Critical' class identification, as reported in our previous chapter 4. The reason was the descriptive nature of the posts along with the narrative of personal experiences (e.g. 'I was beaten', 'He assaulted me'). After the post-processing and stop-words removal, the most common words proved to hold a valuable knowledge about the offenders, victims, and supporters characteristics (e.g. 'he', 'children', 'friends and family'), as well as the duration of the exposure to violence (e.g. 'years'). Following the data collection and 'Gold Standard'

creation, the automatic classification of ATHI mentions was performed. Overall, the ATHI instances comprised the 3.6%, 3% and 2.8% of the datasets (Facebook, Reddit and DV-related blogs respectively). The proportion of ATHI instances was higher, implying greater potential for knowledge discovery about the violent experience in comparison with the associated illnesses or injuries suffered.

In terms of the methodology, two state-of-the-art Deep Learning models were evaluated, namely BLSTMs and stacked residual BLSTMs. To account for the shortcomings of online user-generated content (e.g. misspellings, abbreviations, non-technical expressions) the word embeddings approach was implemented. Word embeddings prove particularly effective with small-to-moderate datasets, by naturally extending the feature sets. The default (GloVe) and alternative (domain-specific) word embeddings were used, following their superior performance in previous applications on similar case studies (i.e. medical embeddings [150], crisis embeddings [251]). The min (d=50) and max (d=300) number of dimensions were further experimented with for the most optimal scenario selection. The main advantage of word-to-vector representation comes from the ability to capture syntactic and semantic relationship between the words. However, the issue may occur with the OOV terms, not present in the training dataset. To address the problem identified, the character embeddings were additionally input into the model training, on top of the word embeddings. Character embeddings aim at exploiting the prefix and suffix information (i.e. 'physical', 'emotional', 'sexual' = Abuse Types, and 'beating' 'battering' = Health Issues) resulting in the classification performance improvement.

In terms of the results, stacked BLSTMs with 300 dimensions yielded the highest accuracy of approx. 78% on average for all three datasets, proving mod-

erately useful in ATHI identification from Social Media streams. Multi-corpus training approach was adopted, where all the instances from distinct corpora were combined to increase the overall volume of the dataset, which in turn improve the classification performance. Thus, the average performance was improved to 80% on average. In terms of the comparison between GloVe and DV embeddings, the difference in accuracy was only negligible, which may not offset the time and cost incurred in domain-specific embeddings development. Still, the words analysis from the DV embeddings enabled to obtain fine-grained insight into the Abuse Types as well as the Health Issues experienced by the victims. In contrast, findings from GloVe proved generic and little informative from DV perspective.

As for the limitations, only moderate data size was used in the experiments due to the expensive process of manual annotation (in particular lengthy and detailed experiences descriptions). Still, the words embeddings applied naturally extends the feature set, effectively leveraging even smaller scale datasets. Further, the state-of-the-art in text classification tasks approaches were adopted, although already existent in literature. Further methodological improvements are considered in future work, based on the experimental results obtained. The benefit of posts extraction from other online platforms (e.g. Twitter, Daily Strength) was recognised as well to further enrich the findings in pro-active support of victims and their families. Despite the limitations identified, the step towards pro-active support and mitigation of the destructive impact of DV on physical and mental health of its victims with state-of-the-art technology has been proposed.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This chapter concludes the thesis by condensing the main contributions of the study performed, including the limitations, future directions, as well as its broader impact within the computational social science area of research.

7.1 Summary of Contributions

DV against women is recognized as an increasing worldwide problem, drawing attention from WHO, and leading to numerous preventative and mitigative strategies development. DV lies at the root of countless societal issues, and due to the sensitive nature of the problem, the topic is often marginalized, preventing its dissolution, and leaving the victims to suffer in silence. As noted in previous research [1, 162], the inhibitions and reservations observed among the abuse victims affect their help-seeking behavior. Due to the various concerns experienced by the victims, ranging from financial security to personal safety, the official support services tend to be under-utilized.

On the other hand, Social Media has emerged as relatively recent phenomenon, and rapidly gained a status of virtual venue, where individuals share meaningful connections, open up about the issues experienced, and form support communities. As a result, Social Media platforms have become a valuable source of knowledge about the problem of violence within the society, as well as the effective channel for the support provision. The involuntarily character of user-generated content, along with an unobtrusive manner of data collection from such platforms have led to the continuously growing popularity of Social

Media-based research for actionable knowledge extraction. Through its worldwide reach and active participation of its users, Social Media has demonstrated an immense potential as a tool for the effective violence prevention and mitigation.

An array of DVCS groups already harness the power of Social Media to educate the communities on the dangers of violence, facilitate the connection between the victims and survivors, and provide the safe space to enable open discussions about the issue. Such initiatives lead to the transformation of victims' lives, bringing the necessary change not only at the individual, but also societal level. However, the current manual approach utilized by the DVCS moderators proves inefficient in handling high-volume of streaming data, affecting the timely responses to the actual instances of abuse. Hence, the thesis set out to answer the primary research question: "How the relevant information from DVCS groups perspective can be identified automatically and accurately in Social Media online discourse with the use of the state-of-the-art technology?." In order to address the question posed, the thesis has focused on developing the advanced classification system to improve the efficiency of DVCS groups in their support of victims of violence. Thus, the objectives of the study are as follows:

- *To develop an approach for automatic detection of critical posts from the irrelevant content;*
- *To introduce a framework for automatic identification of finer-grained classes from critical posts;*
- *To propose a method for automatic extraction of Abuse Types and Health Issues from Personal Story category, previously classified.*

To achieve the objectives set, the thesis has proposed the three distinct benchmark datasets from various Social Media platforms (i.e. Facebook, Reddit and topic-relevant blogs), annotated under the supervision of the domain expert, active in DV domain. In addition to high volume and high velocity, the noisy, informal, descriptive and non-technical character of user-generated content creates further complexity in meaningful knowledge extraction. In such context, the conventional classifiers prove ineffective when exposed to data sparsity, and non-semantic representation. For example, the terms "physical violence", 'physical abuse' and "physical assault" would be treated as separate features, although they share similar meaning in the context of DV. As a result, the classification accuracy would be diminished. Traditional textual features are not scalable nor generalize well given the variety of expressions abound on Social Media. Alternative approaches that are capable of precise syntactic and semantic relationships between the words identification are required to effectively handle the specifics of user-generated content. Therefore, the state-of-the-art Deep Learning with word embeddings has been applied to overcome the shortcomings of traditional methods with respect to the classification performance on high-variety datasets. To the best of our knowledge, this work is the first attempt of Deep Learning empirical validation on the real-world case study of relevant information extraction in the DV domain. Furthermore, to improve the effectiveness of the classification accuracy, the novel DV embeddings were developed. Domain-specific embeddings generation and implementation had demonstrated superior performance in previous studies on online text classification. As a result, approximately 500k posts from a wide range of online sources covering the abuse/violence topics were crawled. These included Facebook, Twitter, Reddit and relevant blogs. Finally, the effectiveness of other fea-

ture extraction techniques commonly adopted in natural language processing tasks such as psycholinguistic (LIWC) and BoW was further evaluated.

Given the problem stated and limitations identified, the Deep Learning based approach for critical posts detection has been proposed **(Chapter 4)**. The content shared on DV dedicated online forums ranges from awareness campaigns, through advice offerings, to empathy expressions. From DVCS groups moderators' perspective, such content is considered 'non-critical', given the lack of indication of abuse instance and urgent support required. On the other hand, in the numerous posts the victims signaled either emotional, financial or informational need from the community (the 'critical' category). Firstly, a benchmark dataset with 'critical' and 'non-critical' labels was constructed. The problem of critical post identification is treated as binary text classification task, where a post is classified based on its textual content. The advanced features such as word embeddings were then extracted from un-structured data in order to account for word-word dependencies. Word embeddings not only accurately capture the linkages between the concepts, but also measure the degree of their similarity, further affecting the classification performance. The pre-trained word embeddings such as GloVe and Word2Vec were used for this purpose. State-of-the-art Deep Learning algorithms (CNNs, RNNs, LSTMs, GRUs and BLSTMs) were subsequently applied, and their resulting accuracies were compared. The highest performing classifier on the prediction task has been identified. Also, the psycholinguistic (LIWC) and textual features (BoW) were further extracted for comparison purposes. The effectiveness of Machine Learning algorithms was additionally explored as a baseline for Deep Learning methods. Empirical evidence on benchmark dataset has demonstrated that Deep Learning models outperformed the conventional approaches. The highest prediction rate of 94%

was achieved by GRUs, while the remaining Deep Learning architectures i.e. CNNs, LSTMs and BLSTMs yielded slightly lower, but comparable results in the experiments conducted. In terms of knowledge discovery, the most distinctive features between the 'critical' and 'non-critical' categories have been investigated for an increased awareness about the violence problem, as reported on Social Media. The analysis of word frequencies highlighted the context i.e. when and where the abuse instances are most likely to take place. The proposed approach proves invaluable for DVCS groups' moderators to immediately identify the critical situation and reach out with the support needed to prevent further harm from happening.

The above mentioned binary classification problem has been subsequently extended to multi-class categorization problem to extract the finer-grained information and assist in the posts' re-routing to specialized services **(Chapter 5)**. The automatic content classification allows the DVCS groups to effectively handle the high-volume and high-velocity data. The 5 classes were identified to evaluate the nature of the problem and further facilitate the DVCS groups requirements needs in the effective support provision. The classes were as follows 'Personal Stories', 'Fund Raising', 'Awareness', 'Empathy' and 'General', and reflected the intent of the user. Similarly to the binary classification, the appropriate benchmark dataset was constructed, considered the 'Gold-Standard'. State-of-the-art Deep Learning models were implemented, and their respective performances were compared. In addition to the pre-trained word embeddings (GloVe and Word2Vec), the domain-specific embeddings were applied. Furthermore, the DV embeddings were evaluated from both the prediction power and knowledge discovery perspective. A comprehensive set of experiments covering all possible feature-model combinations were performed

to identify the most optimal model for multi-class problem. The empirical evidence has demonstrated that Deep Learning models with word embeddings yielded greater performance in comparison with the traditional Machine Learning approaches (except for RNNs). In particular, GRUs with GloVe embeddings achieved the maximum accuracy of approximately 92%. Overall, the classification accuracy of Deep Learning models was marginally higher with DV embeddings and lower number of dimensions (50), proving efficient from computational cost point of view. Further, semantic analysis of the words from the domain-specific embeddings has shown greater potential in meaningful insights extraction about the problem of abuse from directly affected individuals, in comparison with the pre-trained word embeddings. Thus, the fine-grained classes identification proves beneficial in two main aspects (i) it enables the online content to be re-routed to appropriate services such as health care, advocacy organizations or financial support, improving the efficacy of DVCS groups in addressing abusive instances reported on Social Media; (ii) it helps to measure the effectiveness of the awareness promotion campaigns by estimating the public responses to various initiatives, and as result to develop the most effective preventative measures in the future.

Following on the multi-classification study, the novel framework aimed at the Abuse Types and Health Issues extraction from victims' posts was proposed **(Chapter 6)**. Based on the content analysis, the personal stories from the abusive experiences are commonly shared on Social Media platforms. The reasons vary and include advice seeking, interaction need, pain relief etc. The details of violent incidents and associated health impacts before and after exposure to abuse are also disclosed. However, the valuable knowledge about the severity of the issue tend to be buried within the lengthy descriptions of irrelevant

nature. Moreover, the numerous mis-spellings, abbreviations, and ambiguations further complicate the relevant and actionable information extraction. The noisy character of the data as well as variety of the expressions used can be attributed to the diverse cultural background of the victims and their respective linguistic patterns. Hence, the advanced natural language processing and Deep Learning techniques were applied to determine the linkages between the concepts, and identify the most distinctive features automatically, eliminating the reliance on manual feature engineering. The approach further supports scalability, taking into account the popularity of Social Media platforms and ever-growing user-generated content. To automatically extract the Abuse Types and Health Issues from victims' posts, the 3 benchmark datasets from were constructed. The sources such as Facebook, Reddit and relevant blogs were considered. The annotations of the above mentioned Abuse Types and Health Issues were performed from the posts collected. The classes have been defined based on the information needs of policy makers and health-care providers towards improved services design. From the methodology point of view, the 2 popular state-of-the-art models, i.e. BLSTMs and stacked residual BLSTMs, were applied and their performance was compared against the pre-trained and the domain-specific embeddings. The experimental results have shown that the stacked residual BLSTMs model yielded the highest accuracy of approximately 78% for all 3 benchmark datasets. As previously discussed, the DV embeddings have proven efficient in correctly capturing the dependency between the terms, despite the common mis-spellings and abbreviations (e.g. 'physical' as pysical and 'ptsd' as Post Traumatic Stress Disorder). Furthermore, the domain-specific embeddings have provided a valuable insight into the violence severity, as well as the health impact on its victims. Thus, the proposed framework of

DV instances automatic detection is indispensable for DVCS groups to not only pro-actively reach-out to potential victims in time-critical situations, but also to accurately estimate the scale of an issue, facilitating the appropriate measures development.

According to WHO, the various types of information are vital for understanding the circumstances surrounding the violent incidents. The valuable knowledge regarding the illnesses experienced and injuries suffered after an exposure to violence can be effectively mined from the self-reports abound on Social Media. As Social Media is already deeply integrated within our lives, its effective utilization for knowledge discovery about the pressing societal issues has been proposed. In particular, the application of Deep Learning techniques to online discourse on DV-dedicated forums allows to uncover *(i) the circumstances in which abusive relationships developed, (ii) the most common violent incidents occurrence, (iii) the impact on the health of the victims, (iv) the various support services sought-after etc. Thus, the promising approach to leverage the high-volume, high-variety and high-velocity data for community support has been proposed.*

7.2 Study Limitations

Dataset Size: The size of dataset used as part of this study is considered moderate. The main reason is the labour intensive manual labelling due to lack of the publicly available benchmark corpus with fine-grained classes in DV domain. Moreover, the posts extracted from DV-dedicated forums are more descriptive, detailed and lengthy on average, which makes the annotation process even more time-consuming. Nonetheless, the word embeddings have been ap-

plied for features extraction to account for the shortcoming identified. Similar approach is frequently employed when number of posts is limited as word embeddings naturally extend the feature set, diminishing the reliance on large scale datasets.

Dataset Source: There is a certain degree of criticism involved regarding the studies using Social Media data. The main involves the veracity (1 of 5 characteristics of Big Data) of user-generated content, and consequently the reliability of the conclusions arrived at. Social Media frequently can contain incomplete, inaccurate, or even entirely false information. There is a literature referring to memory bias (or cognitive bias) phenomenon stating that humans' perception is subjective, and particular events tend to be exaggerated, neutralized, or even completely erased from memory (such as traumatic ones). The particular type of memory bias - the confirmation bias - occurs when individuals recall the information in a way that asserts one's beliefs. Nonetheless, such instances are considered rare, and form a part of any alternative approach to DV impact estimation, e.g. population-based survey.

Gender Bias: As discussed in Chapter 2, males demonstrate significant inhibitions in reporting the incidents of abuse. Such observation can bias the results to the aspects of violence in which the victim is female. During posts evaluation, the hypothesis that the majority of self-reports is shared by the women was confirmed. Still, the scope of the study covered detection of critical situations from Social Media, and extraction the actionable insight - regardless the gender. Furthermore, the separation of datasets into female-victim/male-victim would result in significant data imbalance, given the female-victim posts domination.

Annotation Errors: Despite the best efforts for precise annotation, as well as the supervision of the domain-expert active in family violence domain, the potential

discrepancies may have occurred. These can result from oftentimes ambiguous, and out-of-context user-generated content. Nonetheless, such risk forms an inherent part of the majority of supervised Machine Learning techniques. Any annotation errors will negatively affect the performance of any automated classifiers, trained on the developed 'Gold Standard'. Still, the Kappa statistic was calculated to ensure the sufficiently high agreement between annotators, and effectively minimize the risk of potential mis-annotations.

Technique Novelty: The objective of this thesis was empirical validation of Deep Learning methods on real-world case study, followed by the knowledge discovery about the DV issue. The state-of-the-art models were applied in order to compare their performance and inform future research. The improvement in accuracy was obtained through the domain-specific embeddings development, integrated into the existing Deep Learning architectures. As a result, the novelty is considered at the practical application to the real-case scenario, for which the gap had been identified. Thus, the major part of the thesis focuses on actionable insights extraction to inform the policy makers, the healthcare providers, and the support groups.

7.3 Future Research Directions

The work can be extended into several promising directions, which can be defined as follows:

New Theories: The findings can be evaluated against the official reports regarding the DV prevalence and severity. The data collected and information extracted as part of this study serves as an alternative source of knowledge about the Abuse Types and Health Issues, to account for the limitations of the existing

approaches. The insights derived can be compared against official reports (if overlapping) for further validation, or new theories development.

Demographics Investigation: The metadata associated with the posts extracted from online forums (Reddit in particular) can serve as a valuable source of information on users' demographics. Such knowledge can further enrich the findings, and lead to interesting insights on the age and gender groups most likely to become either victims, or the abusers. Still, the finer-grained analyses require sufficient amount of data within each sub-group for the results validity.

Geo-location Correlation: The geo-attributes available with Twitter data can be utilized for further study of the Abuse Types and Health Issues patterns across various geographical locations. The spatial information additionally allows for the correlation with regional development indicators (e.g. poverty rate, population density, education level etc.), to uncover potential inter-dependencies, and improve the prediction models.

Dataset Size: The large-scale dataset would be beneficial for the study in terms of the findings validation, as well as the potential for finer-grained analyses performance (e.g. demographics investigation, geo-location correlation). Furthermore, the range of online sources can be extended for wider impact, along with the improved generalisation.

Affected Population: The approach proposed can be applied to study the characteristics of abuse patterns experienced within various sub-populations, e.g. female/male/children/elderly/LGBT. For instance, the male-victim abuse is occasionally discussed in the subreddits such as MensRights, and AskMen. Similarly, to obtain specifically female-victim abuse, the subreddits WomensRights and TwoXChromosomes can be useful.

Classifier Generalisation: Upon retrieval of sufficient amount of data from var-

ious DV-related forums, the more flexible classifier can be developed. Currently, mainly Facebook posts were investigated as the primary source of DV self-reports. The model trained on the multiple sources of data will facilitate the approach generalisation, as well as widen an impact of the research.

Technique Novelty: Despite high accuracies achieved for the state-of-the-art models, further improvement to the techniques used can be proposed. Additionally, the advanced optimisation can be performed, given Deep Learning approach used, which heavily relies on the hyper-parameters selection.

To conclude, the contributions of each section presented in the thesis have been summarised, the study limitations have been shortlisted, and finally the potential future research avenues have been indicated.

BIBLIOGRAPHY

[1] *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence*. Italy, 2013.

[2] abusiverelationships. `https://www.reddit.com/r/abusiverelationships/`.

[3] Adam Acar and Yuya Muraki. Twitter for crisis communication: lessons learned from japan's tsunami disaster. *International Journal of Web Based Communities*, 7(3):392–402, 2011.

[4] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, and M Venkatesan. Twitter sentiment analysis of movie reviews using machine learning techniques. *international Journal of Engineering and Technology*, 7(6):1–7, 2016.

[5] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3906–3918. ACM, 2016.

[6] Chidanand Apté, Fred Damerau, and Sholom M Weiss. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, 12(3):233–251, 1994.

[7] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[8] Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*, 2014.

[9] Womens Health Australia. Partner violence and the health of australian women. 2005.

[10] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, Perth, Australia, 2017.

[11] Sairam Balani and Munmun De Choudhury. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378. ACM, 2015.

[12] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd international conference on computational linguistics: posters*, pages 36–44. Association for Computational Linguistics, 2010.

[13] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. Social sharing of emotions on facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 154–164. ACM, 2015.

[14] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[16] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[17] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[18] Justin C Bosley, Nina W Zhao, Shawndra Hill, Frances S Shofer, David A Asch, Lance B Becker, and Raina M Merchant. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. *Resuscitation*, 84(2):206–212, 2013.

[19] Axel Bruns and Jean E Burgess. # ausvotes: How twitter covered the 2010 australian federal election. *Communication, Politics and Culture*, 44(2):37–56, 2011.

[20] Grégoire Burel, Hassan Saif, and Harith Alani. Semantic wide and deep learning for detecting crisis-information categories on social media. In *International Semantic Web Conference*, pages 138–155. Springer, 2017.

[21] Sandra K Burge, Johanna Becho, Robert L Ferrer, Robert C Wood, Melissa Talamantes, and David A Katerndahl. Safely examining complex dynamics of intimate partner violence. *Families, Systems, & Health*, 32(3):259, 2014.

[22] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[23] Alex Burns and Ben Eltham. Twitter free iran: An evaluation of twitter's role in public diplomacy and information operations in iran's 2009 election crisis. 2009.

[24] The RED HEART Campaign. `https://www.facebook.com/TheREDHEARTCampaign/`, 2015.

[25] Jacquelyn C Campbell. Health consequences of intimate partner violence. *The lancet*, 359(9314):1331–1336, 2002.

[26] Jacquelyn C Campbell, Nancy Glass, Phyllis W Sharps, Kathryn Laughon, and Tina Bloom. Intimate partner homicide: Review and implications of research and policy. *Trauma, Violence, & Abuse*, 8(3):246–269, 2007.

[27] Jacquelyn C Campbell and Jennifer Manganello. Changing public attitudes as a prevention strategy to reduce intimate partner violence. *Journal of Aggression, Maltreatment & Trauma*, 13(3-4):13–39, 2006.

[28] Jacquelyn C Campbell, Daniel Webster, Jane Koziol-McLain, Carolyn Block, Doris Campbell, Mary Ann Curry, Faye Gary, Nancy Glass, Judith McFarlane, Carolyn Sachs, et al. Risk factors for femicide in abusive relationships: Results from a multisite case control study. *American journal of public health*, 93(7):1089–1097, 2003.

[29] Cornelia Caragea, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H Tapia, Lee Giles, Bernard J Jansen, et al. Classifying text messages for the haiti earthquake. In *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011)*. Citeseer, 2011.

[30] Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management*, pages 1–6, 2016.

[31] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

[32] Dave Chaffey. Global social media research summary. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, 2019.

[33] Hao Chen, Susan McKeever, and Sarah Jane Delany. Abusive text detection using neural networks.

[34] Xiaoyi Chen, Carole Faviez, Stéphane Schuck, Lillo-Le Louët, Nathalie Texier, Badisse Dahamna, Charles Huot, Pierre Foulquié, Suzanne Pereira, Vincent Leroux, et al. Mining patients' narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate. *Frontiers in pharmacology*, 9:541, 2018.

[35] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[36] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, 2014.

[37] François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[38] Shaika Chowdhury, Chenwei Zhang, and Philip S Yu. Multi-task pharmacovigilance mining from social media posts. *arXiv preprint arXiv:1801.06294*, 2018.

[39] Freddy Chong Tat Chua, William W Cohen, Justin Betteridge, and Ee-Peng Lim. Community-based classification of noun phrases in twitter. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1702–1706. ACM, 2012.

[40] Rumi Chunara, Jason R Andrews, and John S Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American journal of tropical medicine and hygiene*, 86(1):39–45, 2012.

[41] Rumi Chunara, Jason R Andrews, and John S Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American journal of tropical medicine and hygiene*, 86(1):39–45, 2012.

[42] Jae Eun Chung. Social networking in online support groups for health: how online social networking benefits patients. *Journal of health communication*, 19(6):639–659, 2014.

[43] Rosemary Clark. hope in a hashtag: the discursive activism of# whyistayed. *Feminist Media Studies*, 16(5):788–804, 2016.

[44] Nathan K Cobb, Amanda L Graham, M Justin Byron, Raymond S Niaura, David B Abrams, and Workshop Participants. Online social networks and smoking cessation: a scientific research agenda. *Journal of medical Internet research*, 13(4), 2011.

[45] Nigel Collier and Son Doan. Syndromic classification of twitter messages. In *International Conference on Electronic Healthcare*, pages 186–195. Springer, 2011.

[46] Joshua C Collins. Strategy of career interventions for battered women. *Human Resource Development Review*, 10(3):246–263, 2011.

[47] Price Waterhouse Coopers. A high price to pay: the economic case for preventing violence against women. *report prepared for Our Watch and the Victorian Health Promotion Foundation (VicHealth)*, 2015.

[48] Jaclyn D Cravens, Jason B Whiting, and Rola O Aamar. Why i stayed/left: an analysis of voices of intimate partner violence on social media. *Contemporary Family Therapy*, 37(4):372–385, 2015.

[49] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. Acm, 2010.

[50] Aron Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Language resources and evaluation*, 47(1):217–238, 2013.

[51] Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. Experts and ma-

chines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*, pages 275–281. Springer, 2014.

[52] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *ECIR*, pages 693–696. Springer, 2013.

[53] Yue Dai, Tuomo Kakkonen, and Erkki Sutinen. Minedec: a decision-support model that combines text-mining technologies with two competitive intelligence analysis methods. *International Journal of Computer Information Systems and Industrial Management Applications*, 3(3):165–173, 2011.

[54] DailyStrength. DailyStrength. `https://www.dailystrength.org/`.

[55] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[56] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*, 2014.

[57] Ed De Quincey and Patty Kostkova. Early warning and outbreak detection using social networking websites: The potential of twitter. In *International Conference on Electronic Healthcare*, pages 21–24. Springer, 2009.

[58] Fabio Del Vigna12, Andrea Cimino23, Felice DellOrletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. 2017.

[59] Leon Derczynski, Alan Ritter, Sam Clark, and Kalina 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*, pages 198–206, 2013.

[60] domesticviolence. `https://www.reddit.com/r/domesticviolence/`.

[61] Darren Ellis and John Cromby. Emotional inhibition: A discourse analysis of disclosure. *Psychology & health*, 27(5):515–532, 2012.

[62] Mary Ellsberg, Diana J Arango, Matthew Morton, Floriza Gennari, Sveinung Kiplesund, Manuel Contreras, and Charlotte Watts. Prevention of vi-

olence against women and girls: what does the evidence say? *The Lancet*, 385(9977):1555–1566, 2015.

[63] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[64] Maggie A Evans and Gene S Feder. Help-seeking amongst women survivors of domestic violence: a qualitative study of pathways towards formal and informal support. *Health Expectations*, 19(1):62–73, 2016.

[65] Facebook. Graph API. `https://developers.facebook.com/docs/graph-api`. (Retrieved: 12-06-2017).

[66] Francois Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, 5(Nov):1531–1555, 2004.

[67] Terry Flew, Axel Bruns, Jean Burgess, Orit Ben-Harush, Emma Potter, and Judith Newton. Support frameworks for the use of social media by emergency management organisations. 2015.

[68] George Forman and Evan Kirshenbaum. Extremely fast text feature extraction for classification and indexing. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1221–1230. ACM, 2008.

[69] Michelle Fugate, Leslie Landis, Kim Riordan, Sara Naureckas, and Barbara Engel. Barriers to domestic violence help seeking: Implications for intervention. *Violence against women*, 11(3):290–310, 2005.

[70] Michelle Fugate, Leslie Landis, Kim Riordan, Sara Naureckas, and Barbara Engel. Barriers to domestic violence help seeking: Implications for intervention. *Violence against women*, 11(3):290–310, 2005.

[71] Devin Gaffney. iranelection: Quantifying online activism. In *In Proceedings of the Web Science Conference (WebSci10*. Citeseer, 2010.

[72] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016.

[73] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, 2017.

[74] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

[75] Shang Gao, Michael T Young, John X Qiu, Hong-Jun Yoon, James B Christian, Paul A Fearn, Georgia D Tourassi, and Arvind Ramanthan. Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association*, 25(3):321–330, 2017.

[76] Claudia García-Moreno, Christina Pallitto, Karen Devries, Heidi Stöckl, Charlotte Watts, and Naeema Abrahams. *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence*. World Health Organization, 2013.

[77] Claudia Garcia-Moreno and Charlotte Watts. Violence against women: an urgent public health priority. *Bulletin of the World Health Organization*, 89:2–2, 2011.

[78] Geetika Gautam and Divakar Yadav. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pages 437–442. IEEE, 2014.

[79] Saurav Ghosh, Prithwish Chakraborty, Emily Cohn, John S Brownstein, and Naren Ramakrishnan. Designing domain specific word embeddings: Applications to disease surveillance. *arXiv preprint arXiv:1603.00106*, 2016.

[80] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7:45141, 2017.

[81] Anna Glasier, A Metin Gülmezoglu, George P Schmid, Claudia Garcia Moreno, and Paul FA Van Look. Sexual and reproductive health: a matter of life and death. *The Lancet*, 368(9547):1595–1607, 2006.

[82] Erving Goffman. The presentation of self in everyday life. 1959. *Garden City, NY*, page 259, 2002.

[83] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[84] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013 [in press].

[85] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding (ASRU)*, pages 273–278. IEEE, 2013.

[86] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[87] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[88] Eui-Hong Sam Han, George Karypis, and Vipin Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In *Pacific-asia conference on knowledge discovery and data mining*, pages 53–65. Springer, 2001.

[89] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[90] N Heaivilin, B Gerbert, JE Page, and JL Gibbs. Public health surveillance of dental pain via twitter. *Journal of dental research*, 90(9):1047–1051, 2011.

[91] Marianne Hester. *Making an impact: Children and domestic violence: A reader*. Jessica Kingsley Publishers, 2007.

[92] Stephanie Holt, Helen Buckley, and Sadhbh Whelan. The impact of exposure to domestic violence on children and young people: A review of the literature. *Child abuse & neglect*, 32(8):797–810, 2008.

[93] Shin Hoo-Chang, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn

architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285, 2016.

[94] Jiajia Huang, Min Peng, Hua Wang, Jinli Cao, Wang Gao, and Xiuzhen Zhang. A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web*, 20(2):325–350, 2017.

[95] Qunying Huang and Yu Xiao. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, 4(3):1549–1568, 2015.

[96] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *International Conference on Machine Learning*, pages 754–762, 2014.

[97] Frank Hutter, Jörg Lücke, and Lars Schmidt-Thieme. Beyond manual tuning of hyperparameters. *KI-Künstliche Intelligenz*, 29(4):329–337, 2015.

[98] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 159–162. ACM, 2014.

[99] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. In *ISCRAM*, 2013.

[100] Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. Cross-language domain adaptation for classifying crisis-related short messages. *arXiv preprint arXiv:1602.05388*, 2016.

[101] Nancy E Isaac and V Pualani Enos. *Documenting domestic violence: How health care providers can help victims*. US Department of Justice, Office of Justice Programs, National Institute of , 2001.

[102] Harish Jadhao, Dr Jagannath Aghav, and Anil Vegiraju. Semantic tool for analysing unstructured data. *International Journal of Scientific & Engineering Research*, 3(8), 2012.

[103] Rachel Jewkes. Intimate partner violence: causes and prevention. *The lancet*, 359(9315):1423–1429, 2002.

[104] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[105] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 103–112, Denver, Colorados, 2014.

[106] Michael Jones and Irit Alony. Blogs–the new source of data analysis. 2008.

[107] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[108] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. 2014.

[109] Yoshinobu Kano, William A Baumgartner, Luke McCrohon, Sophia Ananiadou, K Bretonnel Cohen, Lawrence Hunter, and T Tsujii. Data mining: concept and techniques. *Oxford Journal of Bioinformatics*, 25(15):1997–1998, 2009.

[110] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81, 2015.

[111] Kate Kelley, Belinda Clark, Vivienne Brown, and John Sitzia. Good practice in the conduct and reporting of survey research. *International Journal for Quality in health care*, 15(3):261–266, 2003.

[112] Jayashri Khairnar and Mayura Kinikar. Machine learning algorithms for opinion mining and sentiment classification. *International Journal of Scientific and Research Publications*, 3(6):1–6, 2013.

[113] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 174–1751, Doha, Qatar, 2014.

[114] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[115] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[116] Etienne G Krug, James A Mercy, Linda L Dahlberg, and Anthony B Zwi. The world report on violence and health. *The lancet*, 360(9339):1083–1088, 2002.

[117] Anahid Kulwicki, Barbara Aswad, Talita Carmona, and Suha Ballout. Barriers in the utilization of domestic violence services among arab immigrant women: Perceptions of professionals, service providers & community leaders. *Journal of Family Violence*, 25(8):727–735, 2010.

[118] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*, pages 1378–1387, 2016.

[119] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[120] Kathy Lee, Ankit Agrawal, and Alok 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*, pages 1474–1477. ACM, 2013.

[121] Caroline Liou. Using social media for the prevention of violence against women. `http://www.partners4prevention.org/sites/default/files/resources/socialmedia_final.pdf`.

[122] Fasheng Liu and Lu Xiong. Survey on text clustering algorithm. In *2011 IEEE 2nd International Conference on Software Engineering and Service Science*, pages 901–904. IEEE, 2011.

[123] Mingming Liu, Jia Xue, Nan Zhao, Xuefei Wang, Dongdong Jiao, and Tingshao Zhu. Using social media to explore the consequences of domestic violence on mental health. *Journal of interpersonal violence*, page 0886260518757756.

[124] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2873–2879, New York, USA, 2016.

[125] Xiaoqian Liu and Tingshao Zhu. Deep learning for constructing microblog behavior representation to identify social media users personality. *PeerJ Computer Science*, 2:e81, 2016.

[126] TK Logan, Lucy Evans, Erin Stevenson, and Carol E Jordan. Barriers to services for rural and urban survivors of rape. *Journal of interpersonal violence*, 20(5):591–616, 2005.

[127] Sreekanth Madisetty and Maunendra Sankar Desarkar. Aggression detection in social media using deep neural networks. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 120–127, 2018.

[128] Jay A Mancini, John P Nelson, Gary L Bowen, and James A Martin. Preventing intimate partner violence: A community capacity approach. *Journal of Aggression, Maltreatment & Trauma*, 13(3-4):203–227, 2006.

[129] Tony Nicholson Marcia Neave, Patricia Faulkner. Summary and recommendations: Royal commission into family violence. 2016.

[130] Sandra L Martin, Brian Kilgallen, Amy Ong Tsui, Kuhu Maitra, Kaushalendra Kumar Singh, and Lawrence L Kupper. Sexual behaviors and reproductive health outcomes: associations with wife abuse in india. *Jama*, 282(20):1967–1972, 1999.

[131] Marifran Mattson and Jennifer Gibb Hall. *Health as communication nexus: A service-learning approach*. Kendall Hunt Publishing Company, 2011.

[132] Valentina Mazzonello, Salvatore Gaglio, Agnese Augello, and Giovanni Pilato. A study on classification methods applied to sentiment analysis. In *2013 IEEE Seventh International Conference on Semantic Computing*, pages 426–431. IEEE, 2013.

[133] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[134] Heather L McCauley, Amy E Bonomi, Megan K Maas, Katherine W Bogen, and Teagen L O'Malley. # maybehedoesnthityou: Social media underscore the realities of intimate partner violence. *Journal of Women's Health*, 2018.

[135] Judith McFarlane, Lene Symes, John Maddoux, Heidi Gilroy, and Anne Koci. Is length of shelter stay and receipt of a protection order associated with less violence and better functioning for abused women? outcome data 4 months after receiving services. *Journal of interpersonal violence*, 29(15):2748–2774, 2014.

[136] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282, 2012.

[137] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[138] Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, 2016.

[139] Fandong Meng, Zhengdong Lu, Mingxuan Wang, Hang Li, Wenbin Jiang, and Qun Liu. Encoding source language with convolutional neural network for machine translation. *arXiv preprint arXiv:1503.01838*, 2015.

[140] Janet Merrell et al. Social support for victims of domestic violence. *Journal of psychosocial nursing and mental health services*, 39(11):30–35, 2001.

[141] Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.

[142] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceeding of the International Conference on Learning Representations Workshop Track*, pages 1301–3781, Arizona, USA, 2013.

[143] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[144] Terri Beth Miller. The pain and the power: The complex relationship between social media and domestic violence. `https://openforest.net/pain-power-complex-relationship\-social-media-domestic-violence/`, 2017.

[145] Elena Montañés, Javier Fernández, Irene Díaz, Elías F Combarro, and José Ranilla. Measures of rule quality for feature selection in text categorization. In *international Symposium on Intelligent data analysis*, pages 589–598. Springer, 2003.

[146] Jenny Mouzos and Toni Makkai. *Women's experiences of male violence: Findings from the Australian component of the International Violence Against Women Survey (IVAWS)*, volume 56. Australian Institute of Criminology Canberra, 2004.

[147] MS Neethu and R Rajasree. Sentiment analysis in twitter using machine learning techniques. In *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pages 1–5. IEEE, 2013.

[148] Pritam Singh Negi, MMS Rauthan, and HS Dhami. Language model for information retrieval. *International Journal of Computer Applications*, 12(7):13–17, 2010.

[149] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. Sentiment analysis on social media. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 919–926. IEEE, 2012.

[150] Dat Tien Nguyen, Kamla Al-Mannai, Shafiq R Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *ICWSM*, pages 632–635, 2017.

[151] Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. Applications of online deep learning for crisis response using social media information. In *Proceedings of the 4th international workshop on Social Web for Disaster Management*, pages 1–6, Indianapolis, USA, 2016.

[152] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Rapid classification of crisis-related data on social networks using convolutional neural networks. *arXiv preprint arXiv:1608.03902*, 2016.

[153] Thin Nguyen, Thi Duong, Svetha Venkatesh, and Dinh Phung. Autism blogs: Expressed emotion, language styles and concerns in personal and

community settings. *IEEE Transactions on Affective Computing*, 6(3):312–323, 2015.

[154] Thin Nguyen, Bridianne ODea, Mark Larsen, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia Tools and Applications*, 76(8):10653–10676, 2017.

[155] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226, 2014.

[156] Azadeh Nikfarjam, Abeed Sarker, Karen Oconnor, Rachel Ginn, and Graciela 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*, 22(3):671–681, 2015.

[157] R Nivedha and N Sairam. A machine learning based classification for social media messages. *Indian Journal of Science and Technology*, 8(16):1, 2015.

[158] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[159] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[160] Michelle Odlum and Sunmoo Yoon. What can we learn about the ebola outbreak from tweets? *American journal of infection control*, 43(6):563–571, 2015.

[161] World Health Organization. *The world health report 2002: reducing risks, promoting healthy life*. World Health Organization, 2002.

[162] World Health Organization. Understanding and addressing violence against women: Intimate partner violence. Technical report, World Health Organization, 2012.

[163] World Health Organization. *Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence*. World Health Organization, 2013.

[164] World Health Organization. Violence against women: Intimate partner and sexual violence against women - fact sheet, 2017.

[165] World Health Organization et al. Women's mental health: An evidence based review. Technical report, Geneva: World Health Organization, 2000.

[166] World Health Organization et al. Putting women first: ethical and safety recommendations for research on domestic violence against women. Technical report, Geneva: World Health Organization, 2001.

[167] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[168] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[169] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[170] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[171] Bo Pang, Lillian Lee, and Shivakumar 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*, pages 79–86. Association for Computational Linguistics, 2002.

[172] Bo Pang, Lillian Lee, and Shivakumar 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*, pages 79–86. Association for Computational Linguistics, 2002.

[173] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[174] Michael J Paul and Mark Dredze. A model for mining public health topics from twitter. *Health*, 11:16–6, 2012.

[175] Michael J Paul and Mark Dredze. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 168–178, 2013.

[176] Michael J Paul and Mark Dredze. Discovering health topics in social media using topic models. *PloS one*, 9(8):e103408, 2014.

[177] Michael J Paul and Mark Dredze. Discovering health topics in social media using topic models. *PloS one*, 9(8):e103408, 2014.

[178] MJ Paul and M Dredze. A model for mining public health topics from twitter. technical report. *Johns Hopkins University*, 2011.

[179] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8):1226–1238, 2005.

[180] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[181] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar, 2014.

[182] Ruth Petersen, Kathryn E Moracco, Karen M Goldstein, and Kathryn Andersen Clark. Moving beyond disclosure: women's perspectives on barriers and motivators to seeking assistance for intimate partner violence. *Women & health*, 40(3):63–76, 2005.

[183] Saša Petrović, Miles Osborne, and Victor Lavrenko. Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[184] Morgan E PettyJohn, Finneran K Muzzey, Megan K Maas, and Heather L McCauley. # howiwillchange: Engaging men and boys in the# metoo movement. *Psychology of Men & Masculinity*, 2018.

[185] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[186] Nicolai Pogrebnyakov and Edgar Maldonado. Identifying emergency stages in facebook posts of police departments with convolutional and recurrent neural networks and support vector machines. In *5th IEEE International Conference on Big Data. 2017*, pages 4343–4352. IEEE, 2017.

[187] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[188] Martin F Porter. Snowball: A language for stemming algorithms, 2001.

[189] Tania Pouwhare. The effects of family violence on maori womens employment opportunities. *Wellington, National Collective of Independent Womens Refuges Incorporated*, 1999.

[190] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[191] priit kallas. Top 15 Most Popular Social Networking Sites and Apps. `https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/`.

[192] Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*, 2008.

[193] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142, 2003.

[194] G Ramya and PB Sivakumar. Advocacy monitoring of women and children health through social data. *Indian Journal of Science and Technology*, 9(6), 2016.

[195] Susan Rees, Derrick Silove, Tien Chey, Lorraine Ivancic, Zachary Steel, Mark Creamer, Maree Teesson, Richard Bryant, Alexander C McFarlane, Katherine L Mills, et al. Lifetime prevalence of gender-based violence in women and the relationship with mental disorders and psychosocial function. *Jama*, 306(5):513–521, 2011.

[196] Nils Reimers and Iryna Gurevych. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks. *arXiv preprint arXiv:1707.06799*, 2017 [in press].

[197] Harry T Reis and Susan Sprecher. *Encyclopedia of Human Relationships: Vol. 1*. Sage, 2009.

[198] Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. Context-sensitive twitter sentiment classification using neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[199] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE, 2011.

[200] Julian Risch and Ralf Krestel. Aggression identification using deep learning and data augmentation. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 150–158, 2018.

[201] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[202] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[203] Aubrey J Rodriguez, Shannon E Holleran, and Matthias R Mehl. Reading between the lines: The lay assessment of subclinical depression from written self-descriptions. *Journal of Personality*, 78(2):575–598, 2010.

[204] Linda E Rose and Jacquelyn Campbell. The role of social support and family relationships in women's responses to battering. *Health Care for Women International*, 21(1):27–39, 2000.

[205] Arpita Roy, Youngja Park, and SHimei Pan. Learning domain-specific word embeddings from sparse cybersecurity texts. *arXiv preprint arXiv:1709.07470*, 2017.

[206] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

[207] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[208] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[209] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[210] Cicero D Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826, 2014.

[211] NN Sarkar. The impact of intimate partner violence on women's reproductive health and pregnancy outcome. *Journal of Obstetrics and Gynaecology*, 28(3):266–271, 2008.

[212] Abeed Sarker and Graciela Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207, 2015.

[213] Nicolas Schrading, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. # whyistayed,# whyileft: Microblogging to make sense of domestic abuse. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1281–1286, 2015.

[214] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

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

[216] Amit G Shirbhate and Sachin N Deshmukh. Feature extraction for sen-

timent classification on twitter data. *International Journal of Science and Research*, 5(2), 2016.

[217] Le Hoang Son, Hrudaya Kumar Tripathy, Acharya Biswa Ranjan, Raghvendra Kumar, Jyotir Moy Chatterjee, et al. Machine learning on big data: A developmental approach on societal applications. In *Big Data Processing Using Spark in Cloud*, pages 143–165. Springer, 2019.

[218] L Sorensen. User managed trust in social networking-comparing facebook, myspace and linkedin. In *2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology*, pages 427–431. IEEE, 2009.

[219] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[220] Anna Stavrianou, Caroline Brun, Tomi Silander, and Claude Roux. Nlp-based feature extraction for automated tweet classification. *Interactions between Data Mining and Natural Language Processing*, 145, 2014.

[221] Shannon Wiltsey Stirman and James W Pennebaker. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522, 2001.

[222] Carlo Strapparava, Alessandro Valitutti, et al. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Citeseer, 2004.

[223] Sudha Subramani, Hua Wang, Huy Quan Vu, and Gang Li. Domestic violence crisis identification from facebook posts based on deep learning. *IEEE access*, 6:54075–54085, 2018.

[224] Domestic Violence Support. `https://www.facebook.com/domestic.violence.needs.to.stop/`, 2014.

[225] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, Stockholm, Sweden, 2011.

[226] Kateryna M Sylaska and Katie M Edwards. Disclosure of intimate partner

violence to informal social support network members: A review of the literature. *Trauma, Violence, & Abuse*, 15(1):3–21, 2014.

[227] Kateryna M Sylaska and Katie M Edwards. Disclosure of intimate partner violence to informal social support network members: A review of the literature. *Trauma, Violence, & Abuse*, 15(1):3–21, 2014.

[228] Buzhou Tang, Jianglu Hu, Xiaolong Wang, and Qingcai Chen. Recognizing continuous and discontinuous adverse drug reaction mentions from social media using lstm-crf. *Wireless Communications and Mobile Computing*, 2018, 2018.

[229] Break the Silence Against Domestic Violence. `https://www.facebook.com/pg/btsadv/`, 2011.

[230] Quan Tran, Andrew MacKinlay, and Antonio Jimeno Yepes. Named entity recognition with stack residual lstm and trainable bias decoding. *arXiv preprint arXiv:1706.07598*, 2017.

[231] Sabine Trepte, Tobias Dienlin, and Leonard Reinecke. Influence of social support received in online and offline contexts on satisfaction with social support and satisfaction with life: A longitudinal study. *Media Psychology*, 18(1):74–105, 2015.

[232] 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, RGGU*, 2016.

[233] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

[234] Jeanine Warisse Turner, Jean A Grube, and Jennifer Meyers. Developing an optimal match within online communities: An exploration of cmc support communities and traditional support. *Journal of Communication*, 51(2):231–251, 2001.

[235] R Jay Turner and Robyn Lewis Brown. Social support and mental health. *A handbook for the study of mental health: Social contexts, theories, and systems*, 2:200–212, 2010.

[236] Sudha Verma, Sarah Vieweg, William J Corvey, Leysia Palen, James H Martin, Martha Palmer, Aaron Schram, and Kenneth M Anderson. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[237] VicHealth. The health costs of violence: Measuring the burden of disease caused by intimate partner violence. 2004.

[238] VicHealth. The health costs of violence: Measuring the burden of disease caused by intimate partner violence, updated 2014, 2004.

[239] Jessica Vitak and Nicole B Ellison. theresa network out there you might as well tap: Exploring the benefits of and barriers to exchanging informational and support-based resources on facebook. *New Media & Society*, 15(2):243–259, 2013.

[240] Theo Vos, Jill Astbury, LS Piers, A Magnus, M Heenan, Laura Stanley, Laurens Walker, and K Webster. Measuring the impact of intimate partner violence on the health of women in victoria, australia. *Bulletin of the World Health Organization*, 84:739–744, 2006.

[241] Jin Wang, Bo Peng, and Xuejie Zhang. Using a stacked residual lstm model for sentiment intensity prediction. *Neurocomputing*, 322:93–101, 2018.

[242] Sida Wang and Christopher 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*, pages 90–94. Association for Computational Linguistics, 2012.

[243] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[244] Zeerak Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.

[245] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[246] Melinda R Weathers, Jimmy Sanderson, Alex Neal, and Kelly Gramlich. From silence to# whyistayed: Locating our stories and finding our voices. *Qualitative Research Reports in Communication*, 17(1):60–67, 2016.

[247] Xiaocong Wei, Hongfei Lin, Liang Yang, and Yuhai Yu. A convolution-lstm-based deep neural network for cross-domain mooc forum post classification. *Information*, 8(3):92, 2017.

[248] Wikipedia. Reddit. `https://en.wikipedia.org/wiki/Reddit`.

[249] Marsha E Wolf, Uyen Ly, Margaret A Hobart, and Mary A Kernic. Barriers to seeking police help for intimate partner violence. *Journal of family Violence*, 18(2):121–129, 2003.

[250] Umesh Hodeghatta Rao Xavier. Sentiment analysis of hollywood movies on twitter. In *ASONAM*, pages 1401–1404, 2013.

[251] Long Xia, G Alan Wang, and Weiguo Fan. A deep learning based named entity recognition approach for adverse drug events identification and extraction in health social media. In *International Conference on Smart Health*, pages 237–248. Springer, 2017.

[252] Rui Xia and Chengqing Zong. Exploring the use of word relation features for sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1336–1344. Association for Computational Linguistics, 2010.

[253] Jie Yin, Sarvnaz Karimi, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[254] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

[255] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709*, 2017.

[256] Lichi Yuan. Improvement for the automatic part-of-speech tagging based

on hidden markov model. In *2010 2nd International Conference on Signal Processing Systems*, volume 1, pages V1–744. IEEE, 2010.

[257] Ye Yuan and You Zhou. Twitter sentiment analysis with recursive neural networks. *CS224D Course Projects*, 2015.

[258] Jamil Zaki and W Craig Williams. Interpersonal emotion regulation. *Emotion*, 13(5):803, 2013.

[259] Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. Lstm-crf for drug-named entity recognition. *Entropy*, 19(6):283, 2017.

[260] Wei Zhang, Clement Yu, and Weiyi Meng. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 831–840. ACM, 2007.

[261] Xiang Zhang and Yann LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015.

[262] Peng Zhao, Xue Li, and Ke Wang. Feature extraction from micro-blogs for comparison of products and services. In *International Conference on Web Information Systems Engineering*, pages 82–91. Springer, 2013.