IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 14, 2019, accepted March 7, 2019, date of publication April 10, 2019, date of current version April 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908827

Deep Learning for Multi-Class Identification From Domestic Violence Online Posts

SUDHA SUBRAMANI[1], SANDRA MICHALSKA[1], HUA WANG[1], JIAHUA DU[1], YANCHUN ZHANG[1,2], AND HAROON SHAKEEL[3]

[1]Institute for Sustainable Industries and Liveable cities, Victoria University, Melbourne, VIC 8001, Australia
[2]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China
[3]Lahore University of Management Sciences, Lahore 54792, Pakistan

Corresponding author: Sudha Subramani (sudha.subramani1@live.vu.edu.au)

ABSTRACT Domestic violence (DV) is not only a major health and welfare issue but also a violation of human rights. In recent years, domestic violence crisis support (DVCS) groups active on social media have proven indispensable in the support services provided 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, the study aims to: 1) construct the novel ''gold standard' dataset from social media with multi-class annotation; 2) perform the extensive experiments with multiple deep learning architectures; 3) train the domain-specific embeddings for performance improvement and knowledge discovery; and 4) produce the visualizations to facilitate models analysis and results in interpretation. The 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 technology to a real-world problem and proves beneficial to DVCS groups, health care practitioners, and most of all victims.

INDEX TERMS Domestic violence, deep learning, word embeddings, feature extraction, information extraction, knowledge discovery, social media.

I. INTRODUCTION

Domestic Violence (DV) refers to the various acts of abuse such as physical, sexual, emotional or any controlling behaviour within an intimate relationship [1]. DV is one of the most prevailing forms of violence and has become an overwhelming global burden. According to the World Health Organization (WHO) [2], DV has got severe implications on mental and physical health of its victims. To address the problem, WHO [1] has proposed the following preventative and mitigative strategies: (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 for at-risk families. Despite substantial measures undertaken to combat violence against women, factors such as self-blame, disclosure repercussions or fear of being stigmatised [3] prevent the victims from active help-seeking and leads to formal services under-utilisation [4].

The online DVCS groups are promoted for safe advertisement of DV resources, awareness promotion, resource sharing, buddying-up between survivors, non-formal mentoring and fund-raising events, to help the DV victims and their families [4]. The availability of groups such as DVCS on social media platforms (e.g. Facebook, Reddit and Twitter) has encouraged the DV victims to share their personal stories and receive emotional, informational and financial support from te community [5].

However, with the popularity of the support initiatives, the online content generation has grown rapidly in scale, what significantly affected the efficacy of DVCS services with respect to the content prioritisation. Information available on

The associate editor coordinating the review of this manuscript and approving it for publication was Le Hoang Son.

social media, particularly in the early hours of an event has already proven valuable during disaster events [6] and mass emergencies [7]. The extraction of actionable knowledge effectively facilitated the decision makers by providing an instant situational awareness as well as assistance in adequate relief planning [8]. As highlighted above, the time is an important factor in crisis situations management - be it natural catastrophes or domestic abuse acts. Still, the ability to detect critical situations from the deluge of data with no excessive latency has been greatly diminished due to substantial increase in scale of online content generated. The unstructured character of text data has added further complexity to valuable insights extraction.

Deep Learning has already proven successful in text classification tasks, outperforming the benchmark Machine Learning techniques [9]. 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.

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.

In our prior work [10], the approach for *binary classification* of 'critical' versus 'non-critical' online posts using Deep Learning has been proposed. In this paper we present the *multi-class posts categorisation*, providing the finer-grained insight into the violence prevalence and severity from online discourse. 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 violence domain. The categories have been assessed based on their criticality and type of the support needed. For instance, the 'Personal Story' class reports and describes the abusive experience. Thus, its timely detection is of particular importance to DVCS services in order to pro-actively reach out to the victims, before it is too late. Furthermore, the details of violent incidents (physical, emotional etc.) shared by the victims, as well as the related health conditions before/after exposure to abuse (anxiety, depression etc.) effectively assist in public health monitoring in ever-growing online community.

The objectives of the study 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 II provides the background on the problem of violence and the role of social media in its prevention (A), brief explanation of the automatic text classification approach (B), and the successful applications of Deep Learning approach in online text classification tasks. Section III covers methodology followed, namely: data collection from social media (A), 'gold standard' corpora construction (B), features extraction with pre-trained word embeddings (C), Deep Learning models specifications (D), and the performance metrics used (E). Section IV details the experiment design and analysis, including knowledge discovery from the pre-identified classes (A), features extraction and model training (B), the classification accuracies comparison (C), hyper-parameters explanation and evaluation (D), visualisation-supported performance and error analysis (E), and the DV-specific embeddings analysis (F). Section V concludes the results, highlights limitations and proposes future directions for the study.

II. BACKGROUND

A. DOMESTIC VIOLENCE AND SOCIAL MEDIA

DV is one of the most pervasive problems worldwide, and its victims suffer not only from physical, but also sexual, emotional and verbal abuse [11], leading to severe health consequences [12]. DV is one of the leading causes of injuries among women as well [11]. According to Evans and Feder [4], victims are experiencing long waiting times to access specialist healthcare services, and those services are significantly under-utilized. Consider the post *'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 prioritization of content allows DVCS groups to actively reach out to the potential victims, in a timely manner, and with the exact support needed.

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 [13], [14]. Various support services, such as counseling,

crisis hotline, emergency shelter and advocacy services are provided by DVCS [15]. The initiatives established are considered crucial in sufferers mental and physical health improvement [16]. Furthermore, the external support received effectively contributes towards successful violence acts resolution [14]. Still, the sufferers frequently refrain from active help-seeking due to cultural, economic and societal sanctions resulting from leaving the abusive relationship, directly affecting the available support services efficacy [17].

The availability of social media has challenged the notion of violence as private one [18]. According to McCauley *et al.* [19], social media platforms have become the powerful agents for engaging public into a dialogue about the realities of DV [20]–[22]. In terms of the self-disclosure, detailed storytelling, direct and indirect support seeking, and emotional exposure are increasingly observed in virtual environments [23]. As stated by Trepte *et al.* [24], 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 [25] [26]. 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 [27]. 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. The social media campaigns raised against violence play significant role in shaping the openness culture and breaking the silence around the most pressing community issues.

B. AUTOMATIC TEXT CLASSIFICATION

Automatic posts categorization is effectively the classification problem of unstructured textual data. The popular Machine Learning approach to text categorisation have already been applied to tasks such cyberbullying prediction and online harassment [28], [29], emergency situational awareness and crisis response [30], emotion detection and sentiment analysis [31] etc.

Automatic text classification is basically comprised of two main elements, namely *features engineering* and *label prediction*. The first step involves the relevant features extraction from the raw textual data and its numerical vector representation. Some of the widely used features engineering approaches are Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) [32] [33], psycholinguistic features [34], topic modeling features [35]–[38], syntactic relations [39], word n-grams [40], and sentiment lexicon features [41].

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 [42], [43] for text classification tasks are: Support Vector Machine (SVM), Logistic Regression (LR), Decision Trees (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 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 [44]. As an example, consider the terms 'sexual abuse', 'physical abuse' and 'emotional abuse', which represent the various types of abuse 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.

C. APPLICATIONS OF DEEP LEARNING

Deep Learning is the relatively new branch of Machine Learning, with the advantage of automatic features extraction from raw textual data [45]. Deep Learning architectures have already made remarkable improvements in domains such as image processing [46]–[48], pattern recognition and computer vision. The successful applications of Deep Learning have also been observed in Natural Language Processing (NLP) tasks, including Part-of-Speech tagging (POS) [49], sentence modelling [50], machine translation [51], text classification [52], topic categorization [53] etc.

There are two primary Deep Learning models - Convolutional Neural Networks (CNNs) [52] and Recurrent Neural Networks (RNNs) [54]. Both models take the embeddings of words in the text sequence as an input, and generate the real-valued features vectors for those words. CNNs have been applied in the sentence-level sentiment and question classification [50], [52], and proved the advanced performance over traditional Machine Learning techniques such as SVM and MaxEnts [55]. Similarly, RNNs are implemented to model the text sequence achieving an improved performance in multi-class learning [56]. The improved version of RNNs such as Long Short-Term Memory networks (LSTMs) [57], Gated Recurrent Units (GRUs) [58], and

[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*

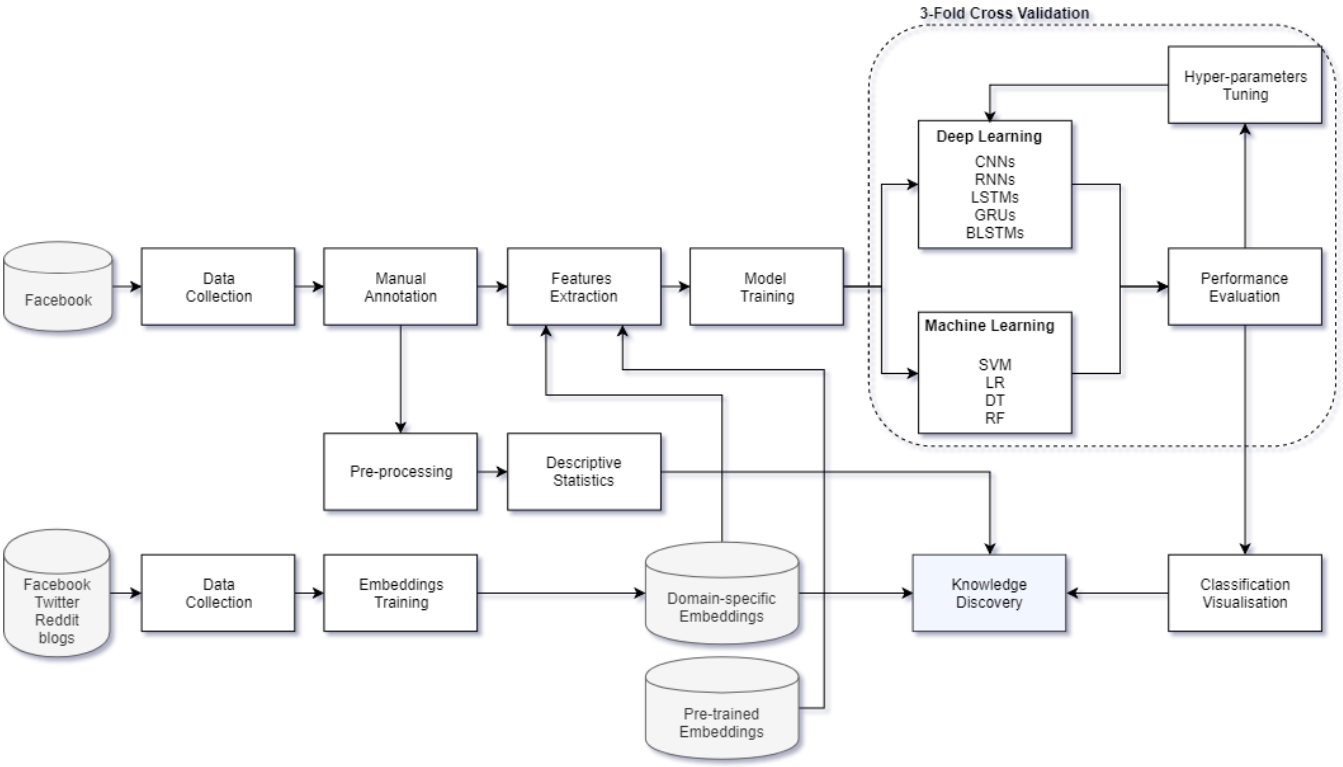


FIGURE 1. **Architecture of Our Proposed Approach for Multi-class Identification.**

Bidirectional LSTMs (BLSTMs) [59] are widely used in NLP 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 [60]. For the similar task conducted in [61], 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 [62], and flood [6]), 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 [63]. For various NLP applications such as sentiment analysis and question-answering [64], GRUs and LSTMs proved superior over the CNNs. In the example of Russian tweets sentiment classification, GRUs achieved higher accuracy than LSTMs and CNNs [65]. 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.

Furthermore, the promising results of Deep Learning are increasingly observed in numerous real-time social media applications. These include abusive language towards racism and sexism detection [61], overtly/covertly/non-aggressive posts prediction [66], crisis information categorisation (e.g advice, donations, infrastructure, sympathy) [62], and DV-critical online posts identification [10]. The study proposed extends our previous work on critical/non-critical posts classification in DV domain by finer-grained details extraction to further support the WHO and DVCS initative raised towards active violence prevention.

III. METHODOLOGY

The section presents the proposed approach (Figure 1) for multi-class DV posts identification from social media. The methodology consists of the five steps, detailed in the following sub-sections.

A. 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 [67]. 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 [68] is that researchers

TABLE 1. Examples of DV Posts and the corresponding information category.

ID	DV Posts	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.'*	*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.'*	*Fund Raising*
P_3	*'Lets stop the violence. Know the signs.'*	*Awareness*
P_4	*'Rest in Peace, Beautiful Angel.'*	*Empathy*
P_5	*'Morning greetings. Enjoy the day.'*	*General*

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.

B. GOLD STANDARD CONSTRUCTION

In order to construct the 'gold standard',[6] 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* - 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 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 credibility 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 [69]. 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.

C. 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

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

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 [70] and Twitter's crawl of GloVe [71] 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 [62]) and named entity recognition (medical embeddings [72]) 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 [70]. 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 [71]. 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 [7] 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, Blogs and Twitter, 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 and 300 embedding dimensions were used for training, given the relatively small size of the dataset in comparison with the pre-trained embeddings.

D. MODEL DEVELOPMENT

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

- **CNNs:** The CNNs architecture used is described in-detail in [53]. 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 [54]. 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 [57], GRUs [58] and BLSTMs [59] 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.

E. 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 [61], [62].

Also, the k-fold cross-validation was applied to assure the robustness of the validation and to prevent overfitting and the potential selection bias [73]. The collected dataset was randomly divided into k partitions, where one partition 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.

IV. 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

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

TABLE 2. 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.

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 Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Decision Trees (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. [74].

(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 mis-classifications. 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.

A. 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 2, 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'.

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.

B. 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 [74]. Additionally, the models were trained up to 50 epochs and implemented using Keras [75].

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 [76]. Accordingly, the dropout rate was set to 0.2 [74].

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 [10], 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.

C. ACCURACY COMPARISON

The dataset was partitioned into training and testing sets, following 3-fold stratified cross-validation approach, as used in previous studies [77], [78]. 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 3.

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 [79]. Given a long sequence, information of initial sequence fades away as the new sequences are fed into the networks of RNNs.

TABLE 3. 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

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 3 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 use the dense representation of words embeddings.

D. 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 2 shows the accuracy of Deep Learning models using both Word2Vec and GloVe word embeddings against various training epochs. The models

TABLE 4. 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 IV-B are underlined in the table.

appear to converge faster with GloVe than Word2Vec features set. With Word2Vec embedding (Figure 2a), the accuracy of Deep Learning models fluctuated at the beginning and became stable after 30 epochs on average. With GloVe embedding (Figure 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 BLSTMs models, as their accuracy scores were the highest. The 3-fold cross

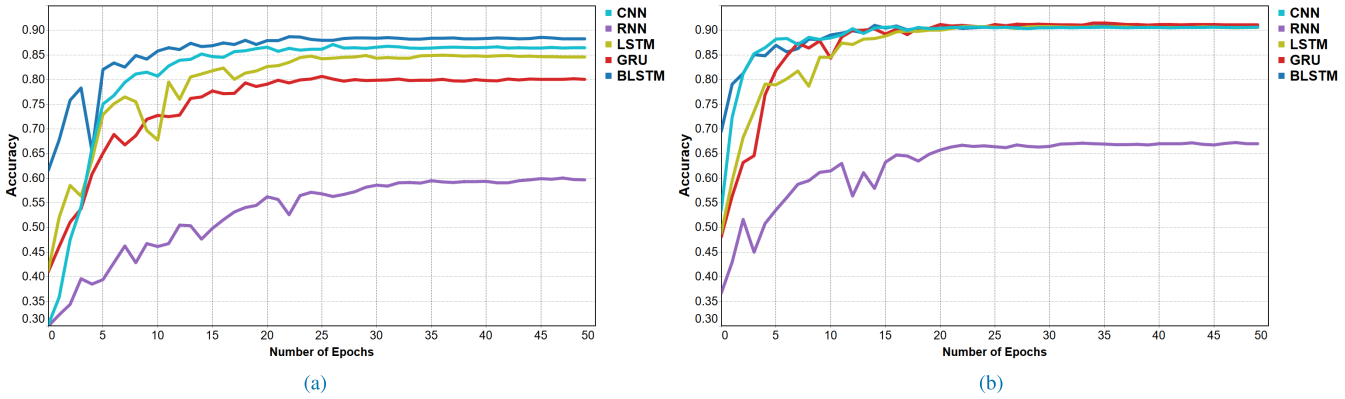


FIGURE 2. Accuracy of Deep Learning Models at Different Epochs. (a) Word2Vec. (b) GloVe.

validation was performed and the outputs are presented in Table 4. 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.

E. 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 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 3. From the analysis, the following conclusions can be drawn:

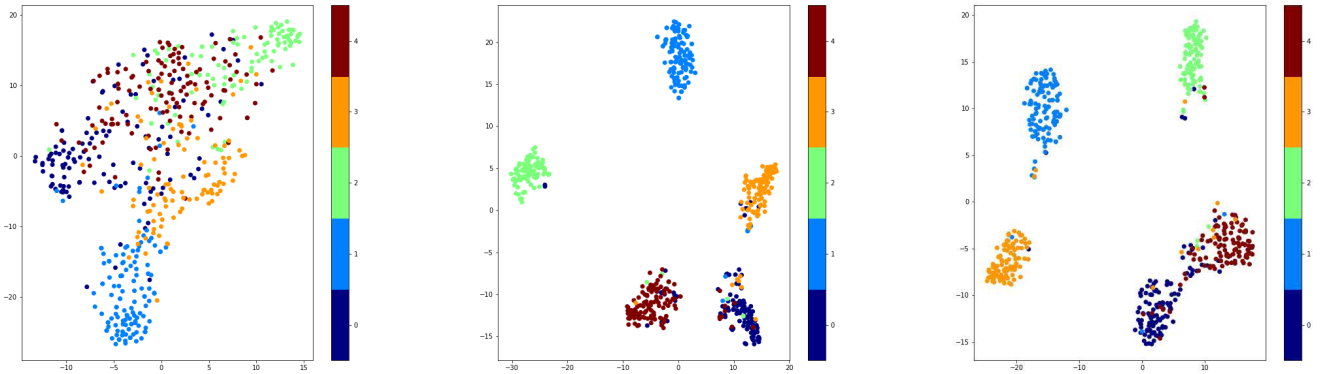
- The posts separation by RNNs (Figure 3a) model did not produce clear distinction between the classes. Additionally, the further overlap for *Fund Raising* and *General* categories occurred. The only clearly segmented group by RNNs model was *Empathy*. It is due to the specific characteristics of posts 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 3b). Minor mis-classifications 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 3c) as well. Nonetheless, the overlap between *Awareness* and *Personal Story* classes was observed in 2D space resulting in higher number of mis-classifications within those groups.

To further validate the findings and quantify the classification accuracy between categories, the confusion matrices were produced for the same models (BLSTMs, GRUs and RNNs) (Figure 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 4a) and GRUs (Figure 4c), respectively. In terms of RNNs (Figure 4a), the group with the most mis-classifications 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.

F. DOMAIN-SPECIFIC EMBEDDINGS ANALYSIS

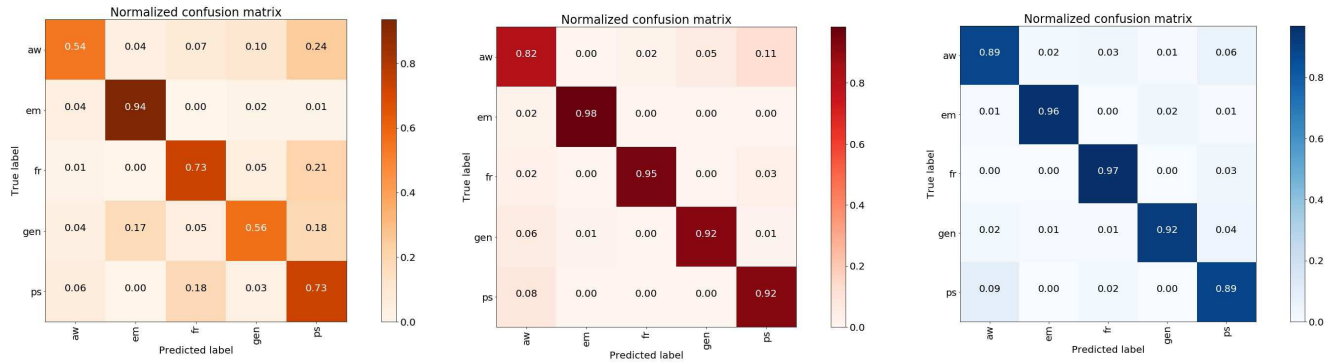
The domain-specific embeddings for DV were constructed and evaluated in terms of (i) impact on the classification performance in Deep Learning models, (ii) useful insight generation and knowledge discovery about the DV. The comparative analysis with GloVe as a benchamark has been performed, and the moderating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

In terms of the average classification accuracy, the difference between GloVe and DV embeddings was negligible,



(a) (b) (c)

FIGURE 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)*. (a) RNNs. (b) BLSTMs. (c) GRUs.



(a) (b) (c)

FIGURE 4. Confusion Matrix of Deep Learning Models w.r.t. GloVe Embedding *(aw-awareness; em-empathy; fr-fund raising; gen-general; ps-personal story)*. (a) RNNs. (b) BLSTMs. (c) GRUs.

TABLE 5. 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**
	BLSTMs	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
	BLSTMs	**91.29**	90.55

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 presents the accuracies obtained for the CNNs, RNNs, LSTMs, GRUs and BLSTMs 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 15 most similar words to the 4 DV-related terms were extracted from both GloVe and DV embeddings (Table 6). The similary was evaluated by the standard measure of cosine distance in the vector space. The 4 DV-related terms, such as 'depression', 'physical', 'abuser' and 'abuse' 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', 'experiencing', '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

TABLE 6. Example words and their top 15 similar words from user posts using DV and GloVe embedding.

DV related Words	Learnt by DV Embedding	Learnt by GloVe Embedding
Depression	*ocd, anxiety, 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*
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*
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*

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

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 violence severity and its impact on the victims.

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', 'batterings'), but also the other abuse types (e.g. 'psychological', 'emotional'). In contrast, GloVe returns mostly irrelevant to DV terms such as 'experience', 'quality' or 'aspects'. Similarly to 'physical', the word 'abuser' and its related terms returned by the domain-specific embeddings allow for deeper insight into, in this case, the types of victimisers, i.e. 'ex', 'partner', 'husband'. Furthermore, the numerous misspellings abound in social media are effectively addressed by DV embeddings through their nearby representation in the vector space (e.g. 'pysical' as 'physical').

From Table 6, the most common potential sources of mis-classifications have been identified and classified as (i) misspellings (e.g. 'violence', 'violenc', 'voilence'), (ii) abbreviations (e.g. 'ocd', 'ptsd', 'dv'), and (iii) synonyms (e.g. 'abuser', 'attacker', 'perpetrator'). The embeddings used for model training effectively address the mis-classifications 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.

V. CONCLUSIONS

Social media has been increasingly used in violence prevention by awareness raising, knowledge sharing, and bringing stories to the public [26]. 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 timely 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 mis-classifications; (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 [80]. 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 3. 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 mis-classifications. As a result, the BLSTMs proved advantageous in the case of *Personal Story* (92% 4b), whereas GRUs in the case of *Fund Raising* and *Awareness* (97% and 89%, respectively 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 [62] and medical embeddings [72]. 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.

REFERENCES

[1] W. H. Organization. (2012). *Understanding and Addressing Violence Against Women: Sexual Violence*. [Online]. Available: http://apps.who.int/iris/bitstream/10665/77433/1/WHO_RHR_12.35_eng.pdf

[2] World Health Organization. (2013). *Global and Regional Estimates of Violence Against Women: Prevalence and Health Effects of Intimate Partner Violence and Non-Partner Sexual Violence*. Italy. [Online]. Available: http://apps.who.int/iris/bitstream/10665/85239/1/9789241564625_eng.pdf

[3] K. R. Henning and L. M. Klesges, "Utilization of counseling and supportive services by female victims of domestic abuse," *Violence Victims*, vol. 17, no. 5, pp. 623–636, 2002.

[4] M. A. Evans and G. S. Feder, "Help-seeking amongst women survivors of domestic violence: A qualitative study of pathways towards formal and informal support," *Health Expectations*, vol. 19, no. 1, pp. 62–73, 2016.

[5] J. E. Chung, "Social networking in online support groups for health: How online social networking benefits patients," *J. Health Commun.*, vol. 19, no. 6, pp. 639–659, 2014.

[6] C. Caragea, A. Silvescu, and A. H. Tapia, "Identifying informative messages in disaster events using convolutional neural networks," in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2016, pp. 1–6.

[7] S. Verma *et al.*, "Natural language processing to the rescue? Extracting 'situational awareness' tweets during mass emergency," in *Proc. ICWSM*, 2011, pp. 1–9.

[8] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting information nuggets from disaster-related messages in social media," in *Proc. ISCRAM*, 2013, pp. 1–10.

[9] D. T. Nguyen, S. Joty, M. Imran, H. Sajjad, and P. Mitra. (2016). "Applications of online deep learning for crisis response using social media information." [Online]. Available: https://arxiv.org/abs/1610.01030

[10] S. Subramani, H. Wang, H. Q. Vu, and G. Li, "Domestic violence crisis identification from facebook posts based on deep learning," *IEEE Access*, vol. 6, p. 54 075–54 085, 2018.

[11] D. Houry *et al.*, "Does screening in the emergency department hurt or help victims of intimate partner violence?" *Ann. Emergency Med.*, vol. 51, no. 4, pp. 433–442, 2008.

[12] L. Heise, M. Ellsberg, and M. Gottmoeller, "A global overview of gender-based violence," *Int. J. Gynecol. Obstetrics*, vol. 78, no. S1, pp. S5–S14, 2002.

[13] K. M. Sylaska and K. M. Edwards, "Disclosure of intimate partner violence to informal social support network members: A review of the literature," *Trauma, Violence, Abuse*, vol. 15, no. 1, pp. 3–21, 2014.

[14] L. E. Rose and J. Campbell, "The role of social support and family relationships in women's responses to battering," *Health Care Women Int.*, vol. 21, no. 1, pp. 27–39, 2000.

[15] L. Bennett, S. Riger, P. Schewe, A. Howard, and S. Wasco, "Effectiveness of hotline, advocacy, counseling, and shelter services for victims of domestic violence: A statewide evaluation," *J. Interpersonal Violence*, vol. 19, no. 7, pp. 815–829, 2004.

[16] B. Liang, L. Goodman, P. Tummala-Narra, and S. Weintraub, "A theoretical framework for understanding help-seeking processes among survivors of intimate partner violence," *Amer. J. Community Psychol.*, vol. 36, nos. 1–2, pp. 71–84, 2005.

[17] A. Kulwicki, B. Aswad, T. Carmona, and S. Ballout, "Barriers in the utilization of domestic violence services among arab immigrant women: Perceptions of professionals, service providers & community leaders," *J. Family Violence*, vol. 25, no. 8, pp. 727–735, 2010.

[18] C. Liou. *Using Social Media for the Prevention of Violence Against Women*. [Online]. Available: http://www.partners4prevention.org/sites/default/files/resources/socialmedia_final.pdf

[19] H. L. McCauley, A. E. Bonomi, M. K. Maas, K. W. Bogen, and T. L. O'Malley, "#MaybeHeDoesntHitYou: Social media underscore the realities of intimate partner violence," *J. Women's Health*, vol. 27, no. 7, pp. 1–7, 2018.

[20] S. Subramani, H. Wang, M. R. Islam, A. Ulhaq, and M. O'Connor, "Child abuse and domestic abuse: Content and feature analysis from social media disclosures," in *Proc. Australas. Database Conf.* Springer, 2018, pp. 174–185.

[21] S. Subramani, H. Q. Vu, and H. Wang, "Intent classification using feature sets for domestic violence discourse on social media," in *Proc. 4th Asia-Pacific World Congr. Comput. Sci. Eng. (APWC CSE)*, Dec. 2017, pp. 129–136.

[22] S. Subramani and M. O'Connor, "Extracting actionable knowledge from domestic violence discourses on social media," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 5, no. 17, p. e2, 2018.

[23] N. Andalibi, O. L. Haimson, M. De Choudhury, and A. Forte, "Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity," in *Proc. 2016 CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: ACM, 2016, pp. 3906–3918.

[24] S. Trepte, T. Dienlin, and L. Reinecke, "Influence of social support received in online and offline contexts on satisfaction with social support and satisfaction with life: A longitudinal study," *Media Psychol.*, vol. 18, no. 1, pp. 74–105, 2015.
[25] M. R. Weathers, J. Sanderson, A. Neal, and K. Gramlich, "From silence to #whyistayed: Locating our stories and finding our voices," *Qualitative Res. Rep. Commun.*, vol. 17, no. 1, pp. 60–67, 2016.
[26] R. Clark, "'Hope in a hashtag': The discursive activism of #whyistayed," *Feminist Media Stud.*, vol. 16, no. 5, pp. 788–804, 2016.
[27] M. E. PettyJohn, F. K. Muzzey, M. K. Maas, and H. L. McCauley, "#HowIWillChange: Engaging men and boys in the #MeToo movement.," in *Psychology of Men & Masculinity*, 2018.
[28] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA)*, vol. 2, Dec. 2011, pp. 241–244.
[29] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Proc. ECIR*. Springer, 2013, pp. 693–696.
[30] M. A. Cameron, R. Power, B. Robinson, and J. Yin, "Emergency situation awareness from twitter for crisis management," in *Proc. 21st Int. Conf. World Wide Web*. New York, NY, USA: ACM, 2012, pp. 695–698.
[31] M. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in *Proc. 4th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2013, pp. 1–5.
[32] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.
[33] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.
[34] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.
[35] M. Peng *et al.*, "Topic-net conversation model," in *Proc. Int. Conf. Web Inf. Syst. Eng.* Springer, 2018, pp. 483–496.
[36] M. Peng, Q. Xie, H. Wang, Y. Zhang, and G. Tian, "Bayesian sparse topical coding," *IEEE Trans. Knowl. Data Eng.*, to be published.
[37] M. Peng *et al.*, "Mining event-oriented topics in microblog stream with unsupervised multi-view hierarchical embedding," *ACM Trans. Knowl. Discovery from Data*, vol. 12, no. 3, p. 38, 2018.
[38] M. Peng *et al.*, "Neural sparse topical coding," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2332–2340.
[39] R. Xia and C. Zong, "Exploring the use of word relation features for sentiment classification," in *Proc. 23rd Int. Conf. Comput. Linguistics, Posters*. Assoc. Comput. Linguistics, 2010, pp. 1336–1344.
[40] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics,*. vol. 2, Assoc. Comput. Linguistics, 2012, pp. 90–94.
[41] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *J. Artif. Intell. Res.*, vol. 50, pp. 723–762, Aug. 2014.
[42] S. Subramani, S. Michalska, H. Wang, F. Whittaker, and B. Heyward, "Text mining and real-time analytics of twitter data: A case study of australian hay fever prediction," in *Health Inf. Sci. 7th Int. Conf. Health Inf. Sci.*, Cairns, QLD, Australia, Oct. 2018, pp. 134–145.
[43] L. H. Son *et al.*, "Machine learning on big data: A developmental approach on societal applications," in *Big Data Processing Using Spark in Cloud*. Springer, 2019, pp. 143–165.
[44] Z. Xu, M. Chen, K. Q. Weinberger, and F. Sha, "An alternative text representation to TF-IDF and bag-of-words," *CoRR*, Jan. 2013.
[45] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[46] D. J. Hemanth, J. Anitha, A. Naaji, O. Geman, D. E. Popescu, and L. H. Son, "A modified deep convolutional neural network for abnormal brain image classification," *IEEE Access*, vol. 7, pp. 4275–4283, 2018.
[47] D. J. Hemanth, J. Anitha, and L. H. Son, "Brain signal based human emotion analysis by circular back propagation and deep Kohonen neural networks," *Comput. Elect. Eng.*, vol. 68, pp. 170–180, May 2018.
[48] C. N. Giap, L. H. Son, and F. Chiclana, "Dynamic structural neural network," *J. Intell. Fuzzy Syst.*, to be published.
[49] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, Aug. 2011.
[50] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," Tech. Rep., 2014.
[51] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," Tech. Rep., 2016.
[52] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, 2014, pp. 1–6.
[53] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *Human Lang. Technol., Annu. Conf. North Amer. Chapter ACL*, Denver, Colorados, 2014, pp. 103–112.
[54] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2011, pp. 1017–1024.
[55] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *Proc. IJCAI-Workshop Mach. Learn. Inf. Filtering*, vol. 1, 1999, pp. 61–67.
[56] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, New York, NY, USA, 2016, pp. 2873–2879.
[57] A. Graves. (2013). "Generating sequences with recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1308.0850
[58] K. Cho and B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, Doha, Qatar, 2014, pp. 103–111.
[59] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, 2013, pp. 273–278.
[60] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proc. 1st Workshop Abusive Lang. Online*, 2017, pp. 85–90.
[61] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, Perth, WA, Australia, 2017, pp. 759–760.
[62] D. T. Nguyen, K. Al-Mannai, S. R. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," in *Proc. ICWSM*, 2017, pp. 632–635.
[63] N. Pogrebnyakov and E. Maldonado, "Identifying emergency stages in facebook posts of police departments with convolutional and recurrent neural networks and support vector machines," in *Proc. 5th IEEE Int. Conf. Big Data.*, Dec. 2017, pp. 4343–4352.
[64] W. Yin, K. Kann, M. Yu, and H. Schütze0. (2017). "Comparative study of cnn and rnn for natural language processing." [Online]. Available: https://arxiv.org/abs/1702.01923
[65] J. Trofimovich, "Comparison of neural network architectures for sentiment analysis of russian tweets," in *Proc. Comput. Linguistics Intellectual Technol., Int. Conf. Dialogue (RGGU)*, 2016, pp. 1–10.
[66] J. Risch and R. Krestel, "Aggression identification using deep learning and data augmentation," in *Proc. 1st Workshop Trolling, Aggression Cyberbullying (TRAC)*, 2018, pp. 150–158.
[67] k. priit. (Aug. 16, 2018). *Top 15 Most Popular Social Networking Sites and Apps*. [Online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/
[68] Facebook. (Jun. 12, 2017). *Graph API*. [Online]. Available: https://developers.facebook.com/docs/graph-api
[69] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Med.*, vol. 22, no. 3, pp. 276–282, 2012.
[70] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent. Workshop Track*, Arizona, AZ, USA, 2013, pp. 1–12.
[71] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, 2014, pp. 1532–1543.
[72] L. Xia, G. A. Wang, and W. Fan, "A deep learning based named entity recognition approach for adverse drug events identification and extraction in health social media," in *Proc. Int. Conf. Smart Health*. Springer, 2017, pp. 237–248.
[73] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.
[74] N. Reimers and I. Gurevych. (2017). "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks." [Online]. Available: https://arxiv.org/abs/1707.06799
[75] F. Chollet. (Jan. 25, 2018). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[76] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1019–1027.
[77] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, vol. 10, Assoc. Comput. Linguistics, 2002, pp. 79–86.
[78] S. Hoo-Chang *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[79] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[80] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *J. Biomed. Informat.*, vol. 53, pp. 196–207, Feb. 2015.

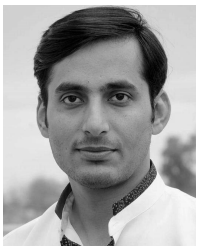
SUDHA SUBRAMANI received the B.E and M.E degrees in computer science from Anna University, India, in 2010 and 2012, respectively. She is currently pursuing the Ph.D. degree with the Centre for Applied Informatics, College of Engineering and Science, Victoria University, Australia, and is also a Research Ambassador. She has done summer internship in Robert Bosch, Japan, from 2016 to 2017, and worked on a project to develop disease assistance tool for disease management in tomato green houses. Her research interests include social media data analytics, data mining, machine learning, deep learning, and text mining. She received the Gold Medal for her academic excellence in the master degree. She was a recipient of the International Postgraduate Research Scholarship.

SANDRA MICHALSKA received the B.Sc. degree in production engineering and the M.Sc. degree in management and information systems. She is currently pursuing the Ph.D. degree in computer science with Victoria University. Her research involves the development of cognitive computing platforms for business applications. In particular, the application of machine learning techniques (recently deep learning) to natural language processing and image recognition tasks. She has working experience as an Intelligence Analyst in public sector in U.K. She is also a member of the British Computer Society with certifications in Business Analysis Practice, Business Processes Modelling, Requirements Engineering, and Commercial Intelligence.

HUA WANG received the Ph.D. degree from the University of Southern Queensland, Australia. He was a Professor with the University of Southern Queensland before he joined Victoria University. He is currently a full time Professor with Victoria University. He has more than ten years teaching and working experience in applied informatics at both enterprise and university. He has expertise in electronic commerce, business process modeling, and enterprise architecture. As a Chief Investigator, three Australian Research Council (ARC) Discovery grants have been awarded, since 2006, and 200 peer-reviewed scholar papers have been published. Six Ph.D. students have already graduated under his principal supervision.

JIAHUA DU received the B.Sc. and M.Sc. degrees in computer science from South China Normal University, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree in computer science with Victoria University. His research interests include natural language processing, data mining, machine learning, deep learning, and social media data analytics. He is currently working on a project regarding helpfulness detection and knowledge discovery from electronic word-of-mouth.

YANCHUN ZHANG received the Ph.D. degree in computer science from The University of Queensland, in 1991. He is currently the Director of the Centre for Applied Informatics, Victoria University, and coordinates a multidisciplinary e-research program across the University. His research interests include databases, data mining, web services, and e-health. He has published over 260 research papers in international journals and conference proceedings including top journals such as *ACM Transactions on Computer-Human Interaction* (TOCHI), the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), and a dozen of books and journal special issues in the related areas. He is the Chairman of the International Web information Systems Engineering Society (WISE). He is a Founding Editor and Editor-In-Chief of *World Wide Web* Journal (Springer) and *Health Information Science and Systems* Journal (BioMed Central), and also the Founding Editor of the *Web Information Systems Engineering* Book Series and *Health Information Science* Book Series.



HAROON SHAKEEL joined the Lahore University of Management Sciences (LUMS), Lahore, Pakistan, in 2015, as a Ph.D. Scholar. He is working in natural language processing and deep learning. He is also the Team Lead with the Knowledge and Data Engineering (KADE) lab, Computer Science Department, LUMS.