Deep Learning for Multi-Class Antisocial Behavior Identification From Twitter

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 16, 2020, accepted October 2, 2020, date of publication October 13, 2020, date of current version November 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3030621

Deep Learning for Multi-Class Antisocial Behavior Identification From Twitter

RAVINDER SINGH[1], SUDHA SUBRAMANI[1], JIAHUA DU[1], YANCHUN ZHANG[1], (Member, IEEE), HUA WANG[1], (Member, IEEE), KHANDAKAR AHMED[1], (Member, IEEE), AND ZHENXIANG CHEN[2], (Member, IEEE)

[1]Institute for Sustainable Industries and Liveable Cities, Victoria University, Footscray, VIC 3011, Australia
[2]School of Information Science and Engineering, University of Jinan, Jinan 250022, China

Corresponding author: Ravinder Singh (ravinder.singh@vu.edu.au)

This work was supported by the Australian Government Department of Education Skills and Employment.

ABSTRACT Social Media has become an integral part of our daily life. Not only it enables collaboration and flow of information but has also become an imperative tool for businesses and governments around the world. All this makes a compelling case for everyone to be on some sort of online social media platform. However, this virtuousness is overshadowed by some of its shortcomings. The manifestation of antisocial behaviour online is a growing concern that hinders participation and cultivates numerous social problems. Antisocial behaviour exists in its various forms such as aggression, disregard for safety, lack of remorse, unlawful behaviour, etc. The paper introduces a deep learning-based approach to detect and classify online antisocial behaviour (ASB). The automatic content classification addresses the issue of scalability, which is imperative when dealing with online platforms. A benchmark dataset was created with multi-class annotation under the supervision of a domain expert. Extensive experiments were conducted with multiple deep learning algorithms and their superior results were validated against the results from the traditional machine learning algorithms. Visually enhanced interpretation of the classification process is presented for model and error analyses. Accuracy of up to 99% in class identification was achieved on the ground truth dataset for empirical validation. The study is an evidence of how the cutting-edge deep learning technology can be utilized to solve a real-world problem of curtailing antisocial behaviour, which is a public health threat and a social problem.

INDEX TERMS Online antisocial behavior, classification, deep learning, feature extraction, knowledge discovery, information extraction, social media behavior.

I. INTRODUCTION

A personality disorder is an enduring pattern of inner experience and behaviour that deviates markedly from the expectations of an individual's culture, is pervasive and inflexible, has an onset in adolescence or early adulthood, is stable over time, and leads to distress or impairment [1]. There are a total of ten personality disorders and these are grouped into three different clusters. Antisocial behaviour falls in the Cluster B of personality disorders along with borderline, histrionic, and narcissistic personality disorder [1]. Individuals who experience symptoms of these disorders often appear emotional, erratic, and dramatic. Antisocial behaviour is a mental health disorder that has been made popular by movies and television and there is a lot of misunderstanding and misinformation prevailing about it among the general public. There are a lot of different criteria that one can meet, to be classified as to have excerted antisocial behaviour. Some of the characteristics of ASB are the repeated acts that violate our social norms, deceitfulness and lying, impulsivity, irritability and aggressiveness, reckless disregard for the safety of self and others, consistent irresponsibility, and lack of remorse. Irresponsibility could be over one dimension, such as work and family, or across multiple dimensions. An individual with antisocial behaviour personality, when committing an act that harms other people, does not feel guilt or exhibits remorse. A lot of the time a perpetrator tends to blame a victim or imply that the victim deserved to be treated that way, displaying a lack of empathy [2].

A number of people with antisocial personality commit severe crimes, however, that is not the only criterion for someone to have exhibited antisocial behaviour. Just being rude and using taboo words can sometimes be qualified as an antisocial behaviour. ASB is extremely difficult to treat, and

The associate editor coordinating the review of this manuscript and approving it for publication was Jun Wang.

the charming demeanour and the manipulation techniques embraced by offenders makes it even harder to deter it. There are a few behaviour characteristics that are often possessed by offenders and these are lack of empathy, superficial charm, and inflated self-appraisal. Online antisocial behaviour is the manifestation of antisocial behaviour on social media, blogs, news channels, and various other online platforms, that are primarilarly used to express views and to share information. Individuals with antisocial personality, when using these channels, often display disregard for other participants & law, use abusive & threatening language and behave in a socially unaccepted manner. There hasn't been much work done on deterring such behaviour online. Some platforms intentionally let such behaviour to prevail in the name of freedom of speech, however, there is a fine line between the freedom of speech and an act unacceptable social behaviour.

To confront antisocial behaviour online, it is imperative to understand its etiology. There are many factors that can lead to a person developing and manifesting ASB. Some of these factors are parental-rejection, maternal-depression, physical neglect, genetic-influence, and poor nutrition intake, etc. All these factors leading to antisocial behaviour can broadly be categorized into three main categories and these are Environmental, Genetic, and Neural. Among the environmental factors that can lead to an individual developing ASB are: exposure to violence, peer influence, family dysfunction, exposure to antisocial behaviour [2] Research has shown that living in a poor neighbourhood, being part of a disadvantaged community, not having a stable job, living in a female-headed household and being dependent on social security are some of the other environmental factors that can trigger the onset of antisocial behaviour in adults [3], [4]. A child that has been raised by biological parents suffering from ASB has a higher probability of developing antisocial behaviour in adulthood [5]. Some studies have shown that if a child of parents suffering from ASB is raised by adopted parents who do not suffer from antisocial behaviour, the chance of this child developing antisocial behaviour is just average. Though genes play a vital role in an individual's developing antisocial behaviour, the influence can be mitigated by changing the individual's environment [6], [7]. Neural factors related to ASB are studied through functional and structural approaches. Whereas functional studies assess the brain's core activities, the structural approach assesses its morphology and these studies together try to comprehend the core neural regions that affect an individual's cognition functions including, the amygdala, frontal cortex, and anterior cingulate cortex [2].

In our prior work [8], we proposed an approach for binary classification of online antisocial behaviour. Based on the content, the posts were classified as either antisocial or general/non-antisocial. In this paper, we are going a step ahead and present a multi-class tweet categorization technique and a fine-grained insight into antisocial behaviour prevailing online. The proposed approach for automatic antisocial behaviour detection and classification is much more efficient than manual investigation and can be implemented at a scale. After careful analysis and under the supervision of a psychologist, we categorized our tweets into five different classes: Four classes for different types of antisocial behaviour and one for general/non-antisocial category. These classes and corresponding labels are presented in Tabel-1. The categories have been identified based on the frequent occurrence of underlying behaviours.

Following are the main objectives of this study:

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

II. BACKGROUND

A. AUTOMATIC TEXT CLASSIFICATION

To detect antisocial behaviour online is basically a text classification research problem that deals with processing and analyzing unstructured text data. The data could be in the form of posts, blogs, comments, and tweets. Natural language processing in itself is a difficult task as it involves dealing with ambiguous text. The same text can have different meaning depending on the context. The whole process becomes even harder when dealing with online text that often includes miss-spellings, not commonly accepted abbreviations, different slangs, and short words. Regardless of the difficulties, researchers have applied different machine learning approaches to emotion and sentiments analysis [9], online harassment and cyberbullying prediction [10]–[12], crises response and emergency situation awareness [13], domestic violence crises prediction [14], etc. Automatic text classification consists of two different procedures and the first of these are feature engineering. Feature engineering is the process of extracting features from the input data and its numerical vector representation. Features are the way we represent our domain knowledge for the classifier. Some of the most commonly used feature engineering techniques are Term Frequency-Inverse document frequency (TF-IDF), Bag-of-Words [15], [16], Topic modelling features [17], Psycholinguistic features [18], Sentiment lexicon features [19], Word n-grams [20], and Word Frequency [21]. The second step in text classification entails label prediction. In this step a machine learning model is first trained on feature extracted and annotated benchmark data set, also known as ground truth dataset. Once trained, the model is then tested on a new unseen dataset and evaluated on numerous performance metrics. The step is repeated using different machine learning models to depict the one with optimal performance.

Subsequently, the most optimal model is then used in research and production. Some of the most widely used algorithms for text classification in machine learning are Logistic regression, Naïve Bayes, Support Vector Machine, Decision Tree, K-Nearest Neighbors, and Random Forest. All these disparate algorithms may suit different sorts of problem sets, the performance of the aforestated classifiers rely heavily on the feature engineering process [22]. The relevance and quality of the feature extracted are directly proportional to the performance of an algorithm. Occasionally, a model trained on a very precise feature extraction process fails to generalize due to overfitting, and this should be avoided at all cost [23].

Humans have an intrinsic and innate ability to under words and their contexts, however, that is not an ability that computers integrally share with us. The widely used feature extraction techniques such as TF-IDF and Bag-of-Words are sometimes not very effective, when dealing with natural language processing, due to the lack of semantic representation of text corpus and inherent over-sparsity issue. To overcome such shortcoming, the relatively new deep learning approach is more appropriate that enables to capture not only the meaning of different words but also their inter dependencies, which leads to a computer understanding meaning and context of a text. Consider the examples of the following short sentences: 'Lack of remorse', 'No regret', 'absolute disregard', and 'completely indifferent'. Though these phrases are related to antisocial behaviour and they all sort of representing an analogous idea, the traditional feature engineering techniques are unable to capture their semantic relationship and representation. Deep learning using feature engineering techniques such as word2vec and GloVe addresses these shortcomings. The techniques also accommodate for misspellings, synonyms, and abbreviations that are prevalent in data collected from social media, leading to significant performance improvement in machine learning text classification problems.

B. APPLICATION OF DEEP LEARNING

The progression of the neural networks was stalled for many years until we experienced deep learning, which is a relatively new phenomenon in machine learning techniques. Deep learning [24] has shown remarkable achievements in domains such as computer vision, pattern recognition, and image processing. The expeditious and brisk progression of the self-driving car industry and enterprise automation can be conveniently credited to deep learning architectures. Natural language processing techniques and research have also been heavily influenced by these deep neural networks. Application of this can be seen in domains such as topic classification, text classification, machine translation, Part-of-Speech tagging, and sentence modelling.

Primarily, there are two deep learning architecture: Recurrent Neural Network [25] and Convolutional Neural Networks [26]. Both these architectures take the word embeddings of text data as inputs and generate feature vectors, which are numerical representation appropriate for manipulation. Convolutional Neural Networks have been applied for question categorization and sentence-level sentiment analysis and have shown superior performance compared to traditional machine learning algorithms such as Support Vector Machine and MaxEnts [26]–[28]. Likewise, Recurrent Neural Networks have been implemented to model text sequence in a corpus and have demonstrated superior performance on multi-class classification [29]. RNNs are either used in their vanilla form or in one of their variants: Long Short-Term Memory networks (LSTMs) [30], Bidirectional LSTM [31], or Gated Recurrent Units [32]. These variants of RNN have been implemented in Natural Language Processing applications and have demonstrated improved performance due to their inbuilt memory architecture to store long-range dependencies and historical information [33].

Convolutional Neural Networks have been utilized in tweet classification problems and have outperformed Linear Regression classifier with very high precision by classifying tweets into hateful and non-hateful rhetoric [34]. In another similar study [35] LSTM outperformed not only the traditional machine learning algorithms SVM and LR but also surpassed the Convolutional Neural Network. CNNs have been used to classifying text into informative and not-so-informative in numerous research studies related to floods [36] and natural disaster [37]to assist crises management and response efficacy. CNNs demonstrated significantly improved performance in these scenarios compared to Random Forest, Linear Regression and Support Vector Machine. In a similar emergency post-classification study, RNN's outperformed Support Vector Machine and Convolutional Neural networks in [38], LSTM outperformed CNN [14], and GRUs outperformed both CNN and RNNs in its vanilla form [39]. For many Natural language processing applications such as question-answering and sentiment analysis [40], LSTMs and GRUs demonstrated better performance over CNNs [41]. In another comparable study pertinent to sentiment classification from tweets, GRUs outperformed CNNs and LSTMs [42]. Nevertheless, all the aforestated deep learning model demonstrated superior performance and yielded better results when compared to traditional machine learning algorithms in disparate text classification problems. Though the performance of all these deep learning models is quite comparable in many of these studies, the decision to use an optimal model is dependent on the nature of an application and the manipulation of hyper-parameters.

Furthermore, deep learning has taken great strides and have shown promising results in various other real-time social-media applications and these include but not limited to crises information detection [37], characterization of mental health conditions [41], aggressive post prediction [43], cyberbullying detection [44], abusive language pertinent to sexism and racism detection [35], fake news detection [45], clickbait detection [46], and domestic violence analogous post categorization [14], etc. This research paper proposes the use of deep learning algorithms to detect antisocial behaviour and to conduct a fine-grained analysis of its various

forms prevailing online. The study also aims to support initiatives to curb online antisocial behaviour and spread related awareness.

C. ANTISOCIAL BEHAVIOR AND SOCIAL MEDIA

Social media communities have the potential to be supportive, punishing, or anywhere in between [47] These communities can not only offer means to work collaboratively but also an abundance of social and entertainment opportunities. Though these can sometimes become breeding grounds for undesirable antisocial behaviour. In the domain of social studies in general, and psychology in particular, antisocial behaviour has been studied and researched extensively, however when it comes to its manifestation online via forums, social media sites, blogs, etc., the topic is still in its nascent form. It is only with the advent of Facebook in 2004, social media became mainstream, followed by Twitter in 2006, and Instagram in 2010. Social media has a far-reaching impact on modern-day life and has been ingrained in our work, social life, entertainment, and other crucial aspects of our daily life. We shop, conduct business, communicate with friends and families and even entertain ourselves on social media. In nutshell, social media has become an integral and inevitable part of modern life and most of us use it for one reason or the other. Apart from enabling modern-day life, social media has also enlivened the type of behaviour online which otherwise is utterly forbidden.

Antisocial behaviour online is a widespread concern that threatens user participation and free discussions in many online communities. In several cases, it can be devastating for victims and deter them from utilizing social media platforms [48]. Online antisocial behaviour appears in diverse forms. Disrespect to lawful behaviour, irritability and aggressiveness, disregards for safety, and lack of remorse are some of the most prevalent forms of antisocial behaviour online [47], [49], [50]. Antisocial behaviour online is an Internet phenomenon of everyday malice, through which culprits seem to have fun at the expense of others' misery and distress [51]. Besides boredom, attention-seeking, malice, revenge, and sadism, motivation to cause anguish is another factor leading to the manifestation of such behaviour online [52]. Imitation effect also has a role to play in the surge of such behaviour on social media. When people see other people displaying a certain behaviour trait, a lot of them will be inclined to do the same, in line with imitation theory, normalizing that behaviour trait [47], [53]. Online antisocial behaviour prevents victims to go about their business lawfully. Most of the platforms still rely on victims to report directly to the platform for it to take appropriate action [54]. Though there is some sort of mechanism in place through which victims can report such behaviour to the underlying platform, most of these cases go unnoticed as victims are often reluctant to report due to scare of retaliation by the culprit [55], [56]. Even though some victims may report such behaviour to a platform, to manually curtail antisocial behaviour online is a laborious and inconceivable endeavour; therefore, an automatic system is required that can work at a scale [8]. In an effort to promote free speech, most online social media platforms fail to curb online antisocial behaviour. Excessive use of these online platforms has also been linked to individuals age 18-27 to display an elevated level of antisocial behaviour, causing distress to others and enjoying at their expense. [51], [52], [57].

Most of the social media platforms have some sort of measures in place to automatically detect pornography, spam, and nudity, nonetheless, antisocial behaviour has not got the attention it deserves, bearing in mind the devastating impact it can have on the victims [58]. These online platforms entice users on a promise to connect them to the rest of the world for ideas and information, however in hindsight, they inadvertently facilitate the spread of antisocial behaviour and putting a large number of users at risk. To detect and classify cyberbullying automatically, using machine learning has been attempted and accomplished by numerous studies [10], [59]–[62]. Similarly, online aggression has also been automatically detected [49], [63]. Trolling, the other prevailing and analogous behaviour to antisocial behaviour has been automatically detected online using multiple machine learning algorithms [53], [64]–[66]. A numerous research studies in the recent past have employed text analysis and natural language processing techniques to automatically detect and classify information from social media related to domestic violence [14], emergency situation awareness [13], and trolling [53], [64], [67], etc., however, none has attempted to automatically detect and classify different forms of online antisocial behaviour to help prevent its proliferation. Bearing in mind the devastating impact that online antisocial behavour can have on the victims, research to find ways to detect and eliminate such behaviour is imperative. As of early 2020, we could not find even a single research study that has attempted to detect and classify different types of antisocial behaviours from online platforms, indicating a gap in the literature. This is a multi-class text categorization problem related to online antisocial behaviour that no other study has undertaken yet, and is the aim of this current research paper. Since deep learning has shown promising results in numerous text classification research problems [39], [68] [44], [69], [70], the technology was extensively experimented with, by using its different variants and parameter settings, to propose a conceptual framework that can be implemented at a scale to detect different types of online antisocial behaviours, and to make online communities safer place for everyone to participate.

III. METHODOLOGY

The high-level conceptual framework is presented in Figure-1, and the detailed steps are discussed in the following sub-sections.

A. DATA EXTRACTION

Data for this paper was collected from Twitter social media platform using Ncaptuer, which is a browser extension. The extension works with Nvivo software tool that is commonly

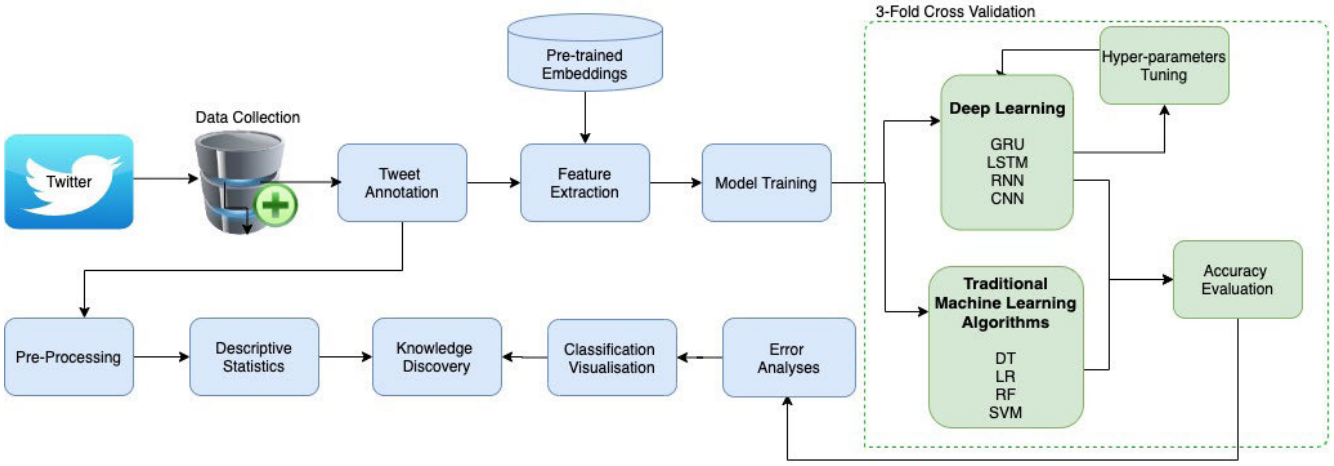


FIGURE 1. Multi-class identification proposed conceptual framework.

used in social science research for qualitative studies. The first step in data collection was to search the Twitter platform for appropriate tweets using pre-decided phrases. Thirty-five such phrases were used to collect the candidate tweets. These phrases included, but not limited to, rude, abusive, and threatening words that are normally associated with online antisocial behaviour. Once the appropriate tweets were discovered, these were collected using Ncapture. Ncaputer saves these tweets in a file with extension.nvcx, which can only be opened in Nvivo, a data analysis software tool. Once opened in Nvivo, the file can then be exported into comma-separated value or an excel file to be used outside of the Nvivo tool to train machine learning and deep learning algorithms. Not all tweets containing rude and threatening words exhibit antisocial behaviour. Therefore, each tweet was manually selected, discarding the ones that did not fit the criteria for antisocial behaviour. The guidelines under DSM-5 [1], which provides diagnostic criteria for behaviour and personality disorders, were observed during the process and it was conducted under the direct supervision of a clinical psychologist.

B. GOLD STANDARD CONSTRUCTION

A gold standard data set was constructed by manually classifying all the tweets that were collected. More than 25000 tweets were initially collected and a lot of these were discarded as these did not meet the required criteria. Since abusive, threatening, and rude language was used in phrases to search for tweets, a lot of tweets returned displayed only sarcasm or were written more as a joke, than to exhibit antisocial behaviour. These were manually filtered out and only those tweets were kept that clearly displayed antisocial behaviour. The tweets that were on borderline were also excluded for training and testing purposes. After filtering and manually annotating, a benchmark corpus of 5504 tweets was created. Tweets were apportioned into five classes each representing a certain category of antisocial behaviour

TABLE 1. Class labels.

Class	Label	Number of Tweets
Failure to conform to social norms	1	1192
Irritability and aggressiveness	2	1238
Reckless disregards for safety	3	804
Lack of remorse	4	868
Non-Antisocial or General Tweets	0	1402

as shown in Table 1. The data set size is considered moderate, given no such study to detected and to classify antisocial behaviour into separate classes has been previously conducted.

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

- *Class 1: Failure to conform to social norms concerning lawful behaviour. It includes individual breaking the law or suggesting breaking it in the future.*
- *Class 2: Irritability and aggressiveness towards others displayed by some sort of aggression.*
- *Class 3: Reckless disregards for the safety of self and others. Not caring for human life (personal or others)*
- *Class 4: Lack of remorse as indicated by being indifferent to or rationalizing having hurt and mistreated.*
- *Class 0: Non-Antisocial or General class that does not exhibit any sort of antisocial behaviour.*

The annotation was performed by the authors with the help of two research students, following the DSM-5 guidelines to classify antisocial behaviour, and under the supervision of a clinical psychologist who specializes in antisocial behaviour disorder. The involvement of a domain specialist was deemed paramount to ensure the accuracy and credibility of the annotation process. To substantiate the inter-rater reliability, the Kappa coefficient, which is one of the most frequently used metrics for similar studies, was calculated. The obtained degree of agreement was 0.91. In case of a discrepancy, the class was assigned by following advice

TABLE 2. Sample classified posts with labels.

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

from the domain specialist. The tweets that were borderline and were not fully matched by the annotators were left out of the study for consistency sake. Some examples of classified tweets and their associated labels are presented in Table 2.

C. FEATURE EXTRACTION

An advantage of using Deep Learning for Natural Language Processing applications in general and multiclass text classification, in particular, is the availability of word embeddings for feature extraction. Word embeddings are text converted into numbers that can be used by deep learning models since these models are unable to process text directly. In technicality, it is the conversion of a text corpus into a feature vector that encapsulate the semantic relationship between words within the corpus, making it is an eloquent representation of text data. In word embedding, the mapping of words takes place in such a way that similar words or similar concepts appear close to each other in vector space, disregarding any misspelling and shortcuts. The abilities of word embeddings to retain the semantic representation of text, automatic feature extraction, and significant dimensionality enables deep learning models to perform superior to the traditional machine learning algorithms and their associated feature extraction techniques, not only by capturing semantic representation but also by overcoming sparsity. For example in the Bag-Of-Word model, the terms 'remorse' and 'regret' are considered as two distinctive features and will be counted separately, however, in word embeddings, their position in the vector space will be very close and will bring in similar semantics to a sentence, as these both imply an analogous concept. All most all the traditional machine learning algorithm use feature engineering techniques such as Word frequency, TF-IDF, and BoW and these classifiers will usually overlook the semantic relationship with similar meaning words, leading to inferior performance compare to word embeddings, when dealing with Natural language processing applications. Most widely used word embedding techniques that were trained on very large external text corpus and have shown tremendous results in various text classification problems are Word2Vec by Google [71], GloVE [72], by Stanford University and FastText [73] by Facebook AI Research.

In this paper, we have used the two most popular and widely accepted word embeddings namely GloVe and Word2Vec. Word2Vec was trained on more than 100 billion words and these were taken from Google News. The words were then mapped to a 300-dimension vector space to construct a vocabulary consisting of 3 million phrases and words [71]. The Glove word embeddings were trained on more than 840 billion words and these words were extracted from Twitter posts. These were then mapped to a 300-dimension vector space to construct a vocabulary of 2.2 million phrases and words [72]. Therefore, in both the word embeddings, every word was mapped to a vector with 300 dimensions.

D. MODEL DEVELOPMENT

In this paper, four different deep learning architectures have been implemented and these are:

- CNNs: The detailed working of Convolutional Neural Networks is described in [74]. The n-grams with most useful information are extracted in the first layer of this model followed by word embedding storage for each word. These are then passed through a pooling layer that produces feature vectors. This convolutional representation is subsequently transformed into an abstract view of a higher level. Lastly, the combination of the composed feature vectors is fed into the dense layer that spits out a corresponding prediction for a text corpus or a post in our case.
- RNNs: The detailed workings of recurrent neural networks are explained in [25]. These networks take an input of variable length sequence using a loop known as the recurrent hidden state. The loop captures information from the previous states of neurons. At every timestamp, a neuron gets an input of information and subsequently update the hidden state. Sentences are just sequences of words and the orders of these words matter to fully understand the contexts and semantics. The structure of sentences and how words are put together in them can convey a comprehensive understanding of semantics; compared to just counting those word individually without any context. RNNs can make use of that ordering and use it as a model, effectively making it well-suited for natural language processing problems.

Therefore, the main advantage of using RNNs over CNNs is that the hidden state within them integrate information from previous time stamps.

- LSTMs and GRUs: GRU [32], LSTM [30], [31] are both enhanced versions of Recurrent neural networks. The fundamental idea behind the LSTMs is their memory units that capture and store historical information over time. Their non-linear gating units regulate the flow of information between neurons and layers. GRUs are essentially the LSTMs, however, unlike LSTMs, which have three gates, GRUs have only two gates. GRUs combine the 'forget' and 'input' gates into one consolidated unit known as the 'update gate. LSTMs are able to integrate contextual information from the previous timestamps enabling the hidden state to capture and utilize this information. Due to these capabilities, both GRUs, and LSTM's are regarded as cutting-edge semantic composition architectures, which are well suited for various text classification problems. These architectures learn and capture long-term semantic and contextual dependencies between words in a text corpus and disregard any information that is redundant.

E. PERFORMANCE EVALUATION

The Accuracy, Precision, F-measure, and Recall are the metrics that were used to evaluate the performance of various classifier used in this research paper. These are the widely accepted evaluation metrics for machine learning algorithms [35], [37]. K-fold cross-validation was also adapted to enforce robustness to the validation process and to impede selection bias and model overfitting, which is a common problem that can occur when trying to improve efficiency by fine-tuning features [75]. In K-fold validation, the data set is randomly split into k number of sets. One of these sets are used for testing and others are for training and validation. The whole process is repeated k times using different training sets, and the results are subsequently averaged to get a final performance metric of an algorithm or a model.

IV. EXPERIMENT DESIGN AND ANALYSIS

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

- *Descriptive Statistics*: Qualitative analyses of the underlying text corpus was conducted in this step. This section presents insight into the types of classes, number of tweets within each class, and the most prevalent words in these classes. Analyses incorporated removing and leaving stop words and implementing stemming.
- *Model Training*: The training procedure for both traditional machine learning and deep learning models is described incorporating the rationale in parameter settings. Both Word2Vec and GloVe models are explored, and feature engineering techniques are presented.
- *Accuracy Evaluation*: The most widely accepted validation metrics i.e. Accuracy, Recall, Precision, and F-measures were calculated and compared, on our benchmark dataset. All the four deep learning architectures, namely RNNs, LSTMs, GRUs, and CNNs were evaluated with these five metrics. For comparison purpose, traditional machine learning algorithms such as RF, SVM, LR, and DT, were also experimented with.
- *Hyper-parameter Evaluation*: Considering an impact the related hyper-parameters can have on the performance of an algorithm or an architecture, a number of experiments were conducted by adjusting various parameter-settings. The parameters that were adjusted for optimization are dropout rate, optimizer type, word embeddings, number of memory units or convolutional filters, and the number of recurrent units. Since training and tuning an artificial neural network can be quite time consuming, the study by Reimers et al [76] was followed for the chosen parameters.
- *Models Visualization*: The confusion matrices and the scatter plots presented illustrate the output performance of all the top performing deep learning architectures implemented. The graphical representation aids in understanding not only the similarities between the different classes but also the misclassifications during the training and testing process. These visualizations present an overview of the classification outputs and assist in understanding the sources of errors. Overall, these are helpful in facilitating the interpretation of model performances.
- *Error Analyses*: The examples of misclassified tweets along with the word embeddings of the commonly occurring terms in the data set are presented in this section. A thorough investigation of the correctly and wrongly classified tweets afford the opportunity for active learning and classification refinement.

A. DESCRIPTIVE STATISTICS

Fine-grained descriptive statistical analyses were performed on the text dataset for comparative purposes and to facilitate knowledge discovery from the different classes. A number of pre-processing tasks were carried out, mainly: (i) Nil pre-processing, (ii) Stop word elimination and (iii) Application of Stemming. The total number of words for each class were counted, along with the maximum and the average number of words in a tweet. Lastly, the most prominent and frequently occurring terms in each class were extracted for the purpose of having a deeper understanding of the nature of disparate classes. The results are shown in the Table-3.

It can be noticed from the table that the total number of words declined significantly after eliminating stop words. In classes 3 and 4 the word count went less than half and in class 0, almost half. This indicates that the generic vocabulary takes up a significant proportion of ASB tweets in all classes.

TABLE 3. Exploratory data analysis of all classes.

Pre-Processing Steps	Word Count	Failure to conform to social norms	Irritability and aggressiveness	Reckless disregard for safety	Lack of remorse	Non Antisocial
Nil	Total Words	21259	23912	23656	22718	22718
	Max. Word	63	65	67	93	61
	Average Word	18	19	29	26	24
	Most Common Words	he, fuck, system, law, i, and, to, a, you, it, my, this, of, is, that, they, in, for, system, her, they, all	you, fuck, i, fucking, the, to, and, a, asshole, bitch, my, of, me, fuckin, it, your, that, is, in, this	i, and, to, the, my, not, a, it, do, on, in, have, me, is, of, safety, with, am, for, that	you, i, and, to, die, have, hope, regret, your, a, that, the, it, in, not, am, do, are, of, for, no	you, i, luck, wish, to, and, the, for, a, in, your, of, it, that, but, with, is, on, my, this
Stop Words Removal	Total Words	12379	13064	10523	9510	17138
	Max. Word	54	41	30	33	53
	Average Word	10	11	13	11	12
	Most Common Words	fuck, system, law, da, em, free, like, get, shit, people, i'm, life, want, fucking, earn, system, legal, make, nigga, pay	fuck, fucking, asshole, bitch, fuckin, cunt, ass, like, whore, shit, get, mother, motherfucker, fucker, say, go, bastard, people	safety, get, almost, like, fun, killed, one, care, fast, life, fire, dangerous, hit, got, know, people, driving, car, head, set	die, hope, love, people, pain, pissing, suffering, deserve, suffer, happy, care, regret, life, know, like, hurt, death, remorse, time, see	luck, wish, pray, like, good, respect, others, know, help, love, great, people, get, i'm, god, one, well, hope
Stemming	Total Words	21259	23912	23656	22718	33115
	Max. Word	63	65	67	93	61
	Average Word	18	19	29	26	24
	Most Common Words	the, fuck, system, law, i, and, to, a, you, it, my, this, that, of, is, they, in, for, system, her, what, they	you, fuck, i, the, to, and, a, bitch, asshol, my, of, me, fuckin, your, it, that, is, in, this, off	i, and, to, the, my, not, a, it, do, on, in, have, me, is, of, safeti, with, am, for, kill	you, i, and, to, die, hope, have, regret, your, a, that, the, it, in, suffer, not, am, do, are, of, for	you, i, luck, wish, to, and, the, for, a, in, your, of, that, it, but, with, is, on, my, this

Word count after stemming remained pretty much the same since no words were eliminated, instead were truncated. Furthermore, from knowledge discovery and interpretation point of view, the application of stemming appeared futile. Due to the nature of the words used in antisocial tweets, a lot of the words did not change much with stemming. Even those that changed such as 'safety' to 'safeti' would not have contributed much to the performance of algorithms because of their lack of meaning.

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

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

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

These are some of the important distinctions in the way people express their behaviours online, and often this is the case in the real world as well, where people use fewer abusive and taboo words to unload their anger, however, feel the need to explain their feelings and disapprovals when showing disregard and lack of remorse for others. The average word count in class 3 and 4 is one and a half times the average word count in class 1 and 2. Since people write lengthier posts when expressing disregard to the safety of others and to display lack of remorse, demonstrates the need for further data mining and knowledge discovery.

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

• *Failure to conform to social norms concerning lawful behaviour:* Apart from the taboo words that are prevalent in the classes 2 (Irritability and Aggressiveness) category

of tweets as well, the most commonly occurring words are related to the law and the legal system. These words are expected in this class since we are dealing with unlawful and illegal behaviour. In addition, the occurrences of 1^{st} and 3^{rd} person pronouns such as 'he', 'her', 'they', 'I', 'my', etc. are predominant in this class. One explanation for this could be that people express their own grudges more towards the legal system than others. Some example are: (1) '*because they took my freedom away'* (2) '*it was not my fault' (3)* '*I was right'*. This demonstrates the importance of some of the stop words in the classification process.

• *Irritability and Aggressiveness:* This class consists of the highest number of aggressive, abusive, taboo, and angry words among all the classes. This is in line with the characteristics of the class. Most prevalent are fuck, fuckin, asshole, bitch, whore, shit, bastard, and motherfucker, etc. It is widely presumed that people often use abusive words to express their aggressiveness and irritability and that is why their prevalence is highest in this category. Presence of 1^{st} and 3^{rd} person pronouns is significantly lower compared to the class 1 tweets.

• *Reckless Disregard for Safety:* The terms in this class significantly deviate from the ones in the last two classes discussed. The use of abusive and taboo words is almost non-existential in this category of tweets. Instead, the antisocial behaviour is expressed using words such as *kill*, *fire*, *dangerous*, *hit*, along with some fun words such as *like*, *fun*, *fast*, etc. People writing these types of tweets seems to have fun at the expense of their own safety, and the safety of others around them. The following tweet sums up the behaviour expressed in this category: "*Woah! Dodged a bullet big time. Ran through red light with no P plate on and just got a warning letter instead of a fine. Thanks NSW Govt!*".

• *Lack of Remorse*: Since this category of tweets relates to the lack of regret after having hurt or mistreated someone, it consists of terms both negative and positive in nature. Negative words such as *suffer*, *die*, *remorse*, *hurt*, *pissing*, *pain* are representatives of having done something wrong or mistreated others, and on contrary, the words such as *love*, *happy*, and *like* may represent the display of indifference and discord after having hurt. An example of a tweet from this class is: "*I'll be laughing when you'll be dying from a curable disease*".

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

B. MODEL TRAINING

The two widely used word embeddings, namely Word2Vec and Glove were used to extract features for deep learning models to gauge and examine their robustness. The first layer in a deep learning model is the embedding layer. By parsing the pre-trained embeddings, this layer executed the index mapping for all the words in the vocabulary and transformed them into dense, fixed-size vectors. The successive layers consisted of 128 memory cells, the number of memory cells commonly used in preceding studies [76]. The models were implemented using Keras [77], a layer built on top of the TensorFlow library from Google [78], and were trained up to 25 epochs to achieve the highest performance.

Unlike traditional machine learning models, in deep learning models, no pre-processing of text was conducted and the whole tweets were fed into the models. In any language, stop words can hold valuable information that can be leveraged to boost model performance. Also, the words in the text were not stemmed. This was avoided to preserve the semantics of each sentence in its original form and to help the models understand the context better. For example, the words *aggression* and *aggressive* can bring in disparate contexts to a text. Initially, the Nadam optimizer was utilized and the batch size was confined to 32 posts with deep learning models, considering the moderate size of our dataset. The number of recurrent units was set to 128 and the activation function used was 'Relu'. The dropout rate was fixed to 0.2 [76]. The 'dropout' is a simple and efficient way to regularize any deep neural network and to prevent overfitting [79]

In regard to the traditional machine learning algorithms, the same Word2Vec and GloVe embeddings were adopted. To overcome the shortcomings of our earlier work [8] i.e. model comparison using simple feature extraction techniques, thorough and comprehensive experimentation was conducted using advanced and widely used feature extraction and model compositions. Python's scikit-learn library with its default parameter settings was implemented for the task of evaluation.

C. ACCURACY EVALUATION

As a part of 3-fold cross-validation approach, the complete dataset was subdivided into training and testing subsets. This approach has been prevalently adopted in numerous precedented studies [80], [81]. The following three pre-processing scenarios for traditional machine learning algorithms were experimented with:

• *Only stemming*
• *Only stop words removal*
• *Both stemming and stop words removal*

The performance of traditional machine learning algorithms depends greatly on the pre-processing steps. The experiments conducted indicated that these algorithms performed best and achieved the highest accuracy on our dataset with stemming only (without removing stop words). In the context of online antisocial behaviour multiclass classification, some stop words could be useful in classes identifications. As discussed in the descriptive statistic section, 1^{st} and 3^{rd} person pronouns were among the frequently occurring words in our tweets, so it made sense to

TABLE 4. Classification models evaluation metrics.

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

keep them in the final dataset as these were part of the context of tweets and assisted in the classification process. The results from 'stemming only' (highest performance) experiments for traditional machine learning algorithms were compared with the results obtained from the four deep learning architectures used, and these are presented in Table-4. Along with the accuracy of these algorithms, other evaluation metrics such as Precision, F-measures, and recall are also presented.

In general, the deep learning architecture's performance was superior to the performance of traditional machine learning algorithms, as indicated by the higher evaluation metrics yield. These algorithms, when used with GloVe embeddings, produced the highest results. RNN's lagged behind in performance using both the GloVe and Word2vec embeddings when compared to the other deep learning and the traditional machine learning algorithms. RNN's inferior performance can be attributed to the difficulty of vanishing gradients [82]. As new sequences during the implementation are fed into the RNNs, information from the preceding sequence diminishes in these architectures. Nonetheless, this limitation or RNNs is apparently addressed by its successors, namely GRUs and LSTMs. These succeeding versions overcome these shortcomings by efficiently capturing long-term dependencies that are imperative when working with textual data, which is sequential in nature.

Both GRUs and LSTMs performed the highest when used with GloVe embeddings and achieved the accuracy of 99.2% and 98.98% respectively. RNNs lagged behind all the other three deep learning architectures and all the traditional machine learning algorithms except the decision tree. When looking into the use of Word2Vec embeddings, GRUs and LSTMs again stood at the top with 98.6% and 98.4% accuracy respectively, and the RNNs again lagged behind. Among the traditional machine learning algorithms, SVM (Support Vector Machine) and LR (Linear Regression) performed the best with 94.99% and 93.36% accuracy respectively, when used with GloVe embeddings. The accuracy was 96.62 and 93.36% respectively when used with Word2Vec. Decision tree's performance was inferior to all the other algorithms used in this study regardless of word embedding combinations. Overall, all algorithms performed better when used with GloVe instead with Word2Vec embeddings, indicating a superior performance capability.

From the results, it can be inferred that the deep learning algorithms have performed better compared to the traditional machine learning algorithms. There is a higher computing cost associated with these algorithms when compared to traditional algorithms, nonetheless, this is compensated with higher performance. The traditional models are suited best for high dimensional and sparse features vectors. It can also be inferred from the results that these are not very well suited for dense vector representation, as used in this study (300 dimensions). The deep learning models can efficiently leverage and use a dense representation of word embedding to attain higher accuracy scores as demonstrated by the results.

D. HYPER-PARAMETER EVALUATION

Deep Learning models performance, used in this study, was evaluated with different hyper-parameter settings and the number of epochs required to get the best possible results. Ideally, too few epochs sometimes can leave a model undertrained and too many epochs can lead to over-fitting. An Underfitted model does not perform well and an overfitted one does not generalize well. To find the right balance is crucial for the best performance of any deep learning model. Another disadvantage of having more epochs than required is the wastage of computing resources. Training deep learning models require a lot of time and computing power and to use anything more than what is required can lead to the wastage of valuable resources. So, experimenting and getting the right number of epochs is paramount. Figure-2 presents comparison of the training process between GloVe and Word2Vec embeddings on our data set. It can be noticed that algorithms achieved their highest accuracy faster when implemented with GloVe compared to Word2Vec.

Comparing the performance when using different optimizers, Nadam and RMSProp produced similar sort of results, with Nadam outperforming slightly because of the

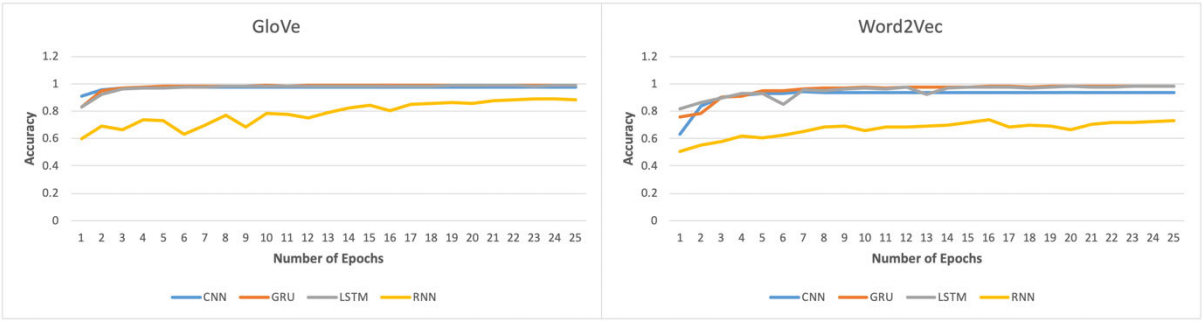


FIGURE 2. Deep Learning Model accuracy and number of epochs.

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

E. MODELS VISUALIZATION

The classification performance of deep learning architectures can be better understood with the aid of visualization. Virtualization provides insights into the inner workings of algorithms. We used t-SNE, the dimensionality reduction techniques, based on GloVe embedding to understand the similarities and dissimilarities among disparate categories of antisocial behaviour. The visualization indicates categories of antisocial behaviour that were correctly classified and the categories that were not. Since algorithms performed better with GloVe embedding compared to Word2Vec, we evaluated visualization generated with Glove embeddings. The highest performing models- GRU and LSTM, and the lowest-performing model-RNN are presented for comparison. The scatter plots in Figure-3 show the clustering of all the five classes. The more confined and distinct the clusters are, the better the algorithm has performed. Following conclusions can be drawn from the analysis of the scatter plots:

• RNN performance on the dataset was relatively inferior when compared with other algorithms used in the study. A significant number of tweet misclassification can be depicted from the scatter plot. The model was unable to generate a clear distinction between some of the classes, especially between the 'General' and the 'Lack of Remorse' categories. The algorithm misclassified a large number of tweets that were meant to be in category 4 (lack of remorse) as non-antisocial tweets. Similarly, it also misclassified some of the category 3 tweets (Reckless disregard for safety) as non-antisocial tweets. It can be inferred that the algorithm was unable to draw a clear distinction between the classes and it is apparent by the lack of sufficient gaps between the clusters representing disparate classes.

• LSTM model performed better than RNN and was able to draw comparatively distinct class clusters. Apart from some misclassifications, clearly defined clusters represent a decent classification performance. As can be seen in Figure-3, some class 4 tweets (Lack of remorse) were wrongly classified as class 3 tweets (Reckless disregard for safety). This is mainly due to the use of similar terms in both types of posts. Phrases such as '*I don't care*', and '*you can die*' were quite common in both classes, and lead to some misclassifications. Similarly, class 2 posts (Irritability and aggressiveness) were wrongly classified as class 1 (Failure to conform to social norms). We believe this may have been again due to the use of similar terms & semantics and sentiments of the underlying posts. The algorithm was able to identify class 1 posts with significantly high accuracy. Overall, the performance was better than RNNs.

• GRU architect was able to distinguish posts fairly correctly, as can be seen in the scatter plot. There is a clearer distinction among classes represented by well-defined clusters with a significant gap between them. Nonetheless, there were a few misclassifications in almost all classes. Some posts from class 1(Failure to conform to social norms) were misclassified as class 3, (reckless disregards for safety). One example is: *'fight the powerfuck the systemkick up a mosh pit when they dont wanna listen'*. Even though the post falls in class 1, the words 'fight' might have led to the algorithm to classify it as class 3 post. Similarly, the following post from class 3 was misclassified as class 4 post, most

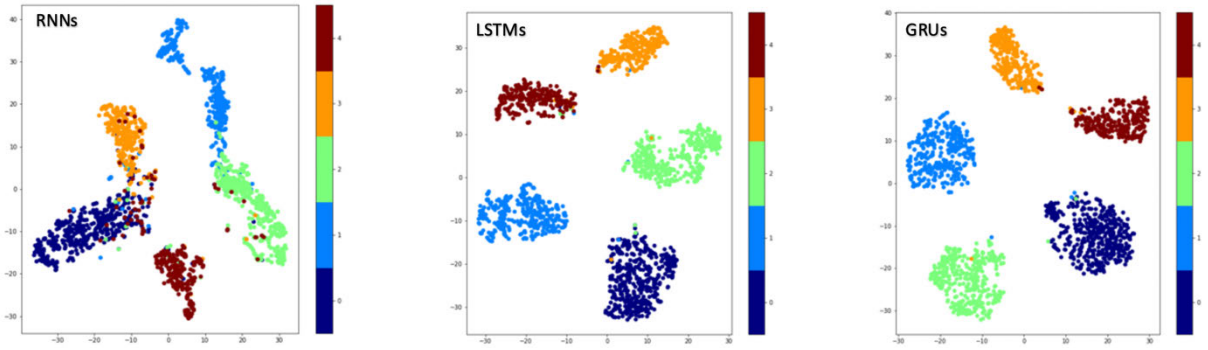


FIGURE 3. Visualization of Antisocial behaviour classes using t-SNE w.r.t GloVe embeddings (0- General/Non-ASB, 1- Failure to conform to lawful behaviour, 2- Irritability and aggressiveness, 3- Reckless disregard for safety, 4- Lack of Remorse).

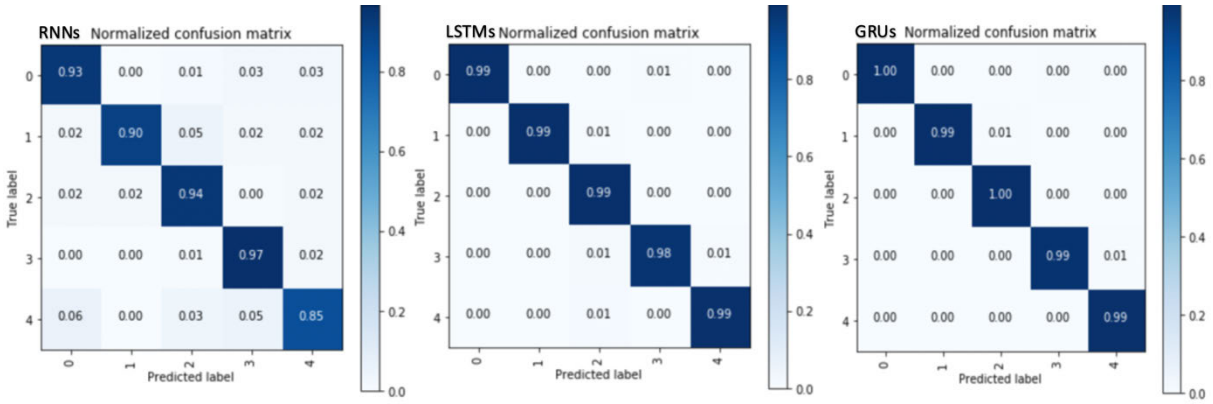


FIGURE 4. Confusion matrix. Deep Learning models w.r.t. GloVe embeddings (0- General/Non-ASB, 1- Failure to conform to lawful behaviour, 2- Irritability and aggressiveness, 3- Reckless disregard for safety, 4- Lack of Remorse).

likely due to the similarity of words in both classes. '*said this so many times and i ll say it till the day i die i do not care for my life i do not care what happens to me i care about what happens to my friends and i care about my friends lives i want everyone of them to succeed and become great'*. There were misclassifications in other four categories as well, however, the performance of GRU architecture was significantly superior to the other algorithms.

To further understand the misclassifications by the architectures that were experimented with, and to quantify the classification accuracies among classes, confusion matrices for the same three architectures (RNN, LSTM, and GRU) were generated and are presented in Figure 4. These confusion matrices provide a finer-grained insight into the classification results. Since 3-fold cross-validation was used in this research, we had three sets of confusion matrices for each architecture, each with its own accuracy and misclassifications. To avoid any interpretation bias, the matrix with the highest accuracy and the lowest misclassifications from the 3-folds were chosen for comparison for all the architectures. RNN's performance was inferior to the other architectures with the highest number of misclassifications. For class 4 (lack of remorse), only 85% of the tweets were correctly classified. 6% were classified as non-antisocial, 5% as class 3 (reckless disregard for safety) and 3% as class 2(Irritability and aggressiveness). Similarly, the algorithm did not perform well in classifying non-antisocial tweets and achieved an accuracy of 90%. 3% were classified as class 3 and another 3% as class 4 tweets. The best it performed was in class 3, with 97% accuracy. In contrast, GRU architecture was able to classify all tweets with high accuracy. It achieved 100% accuracy with class 0 and class 2 tweets. Other three classes were at 99% accuracy each. LSTM performed better than RNN's with fewer misclassifications, nonetheless, did not perform as good as GRUs. Most misclassifications for LSTMs were from class 3 (disregard for safety). It can be inferred that class 2 tweets were mostly classified correctly using all the algorithms and class 4 got most of the misclassifications, indicating that aggression can be depicted relatively correctly compared to lack of remorse.

F. ERROR ANALYSES

The section investigates some of the inaccurate classification outputs; the analysis of which can be leveraged to reannotate some of the tweets and retrain algorithms for further accuracy improvement. The analysis was conducted to understand the source of misclassifications. The examples were generated using Glove-GRU combination that produced the highest results among all the feature-model combinations. Table-5 shows some of the misclassified tweets along with their

TABLE 5. Misclassification examples.

ID	Post	Actual Label	Predicted Label	Probability
P_1	next person that points at my shoes and says what are thoose im responding with go to hell ya	2	1	0.57
P_2	hey anytime fitness stick it up your ass you bunch of crooks this is been cancelled on paper by telephone you re damn right and cancelled by my credit card had to get a new credit card because of you crooks people stay away from anytime fitness scam	2	3	0.45
P_3	we don t need him to go to the cemetery to prove he s a patriot barack obama did every year and he was anything but the patriots put that in your pipe and smoke it or stick it up your ass	2	3	0.54
P_4	what would you do if i jumped from the building cry a bit and then forget	3	4	0.60
P_5	so what if i am speeding on a slipery road on a rainy day its is all fun and part of fun	3	2	0.68
P_6	i do not like wearing protective gears when i am working in the fields i am not worried about my life if will die i will just die no questions	3	4	0.55
P_7	i hope you die slow from the puncture of the blade mais um choker da ogoticario	4	2	0.64
P_8	i love pissing people off with political correctness lmao if you re going to be an arrogant bigot i m going to throw it in your face	4	2	0.53
P_9	i broke her tooth and threw her on the floor not ashamed at all she deserves it	4	3	0.55
P_{10}	2018 10 21 contentment breeds in our disintegration like bitter pills digested by the sick i wish you luck and hope you ve found your medicine	0	2	0.67

actual and predicted label and Table –6 shows the example of correctly classified tweets. Analyzing the output data, it can be discerned that all most all tweets that were classified correctly were classified with a high probability index. However, the tweets that were misclassified, all had a lower probability index. The algorithm had to pick the highest index among the set of all low indexes pointing towards other labels. These tweets that are classified with lower probability can be considered as borderline tweets. There must be some word, phrases or semantic characteristics in these tweets that pointed towards other labels leading the misclassification.

In Table-5, Post-1's (P_1) true label is class-2 (irritability and aggressiveness). The post does show a degree of irritability on part of the writer, however, was classified as class-1 (Failure to conform to social norms). This might be due to the use of word *'hell'*, which is not desirable to be used in this context. This really is a challenging post to classify even by human standards. Similarly, P_4 can also be considered another difficult post to classify due to its context. The post's actual label is class 3 (Reckless disregard for safety), however, it was misclassified as class 4 (Lack of remorse). The post clearly displays disregard for safety when the writer suggests to jump from the building, however, at the same time it exhibits a subtle suggestion that the person to whom the post is referring to, may not show any remorse. There are a few more similar sorts of examples presented in Table-5 and almost all of these are challenging posts to classify due to their confusing structure and word selection. These, along with other misclassified posts can be used to retrain the model as a part of active learning.

In Table-6, the second part of the classification output is presented. These are the tweets that were correctly classified despite containing confusing context and content. The high probability index implies the high degree of confidence on the algorithm's part in classifying these posts. For instance, P_2 was correctly classified as class 1 (Failure to conform to social norms concerning lawful behavior), even though the post has some indication of irritability and aggressiveness (Class-2). Similarly, P_7 was correctly classified as class-4 (lack of remorse), even though the post contains word *'drowning'*, leaning more towards class-3 (Reckless disregards for safety).

Word embeddings output for some of the most commonly occurring and decisive terms in the dataset were generated and are presented in Table-7. In terms of 'regret', mostly similar (e.g. confess, admit) and synonyms (e.g. remorse) words were captured in the vector space, highlighting their interdependence. The word 'broke 'exhibits a more general meaning, and in contrast, attracted a wider range of words implying disparate meanings and contexts. Similar to the word 'regret' the word 'bastard' attracted related and synonyms taboo words. A few more examples are presented, and these afford the opportunity to have a deeper dive into the inner workings of the word embeddings. Notwithstanding the syntactical linguistic variety that is inherited to any social media platform, the deep learning model, utilizing word embeddings, exhibits its ability to distinguish relations between concepts that are imperative to any natural language processing task. The wrongly classified tweets and their corresponding prediction probabilities are returned, that enables the identification of source classification confusion, leading to potential refinement of the classes. Moreover, the analysis of the embeddings facilitates an opportunity for continuous performance improvement and active learning.

The model demonstrates its robustness in gauging the subtle clues within the data even with a relatively small training data set (approx. 5500). Absence of substantial research work on behaviour studies from social platforms makes this work a valued starting point in detecting and eliminating online antisocial behavior. The need for manual feature engineering efforts is eliminated due to the use of deep learning architecture that facilitates a systematic and automated approach. Considering the extremely noisy characteristics of data collected from social media platforms,

TABLE 6. Correctly classified tweet examples.

ID	Post	Actual Label	Predicted Label	Probability
P_1	legally fuck the law laws are merely words on paper backed by the promise of violence if they are not obeyed words on paper have no authority over any of us just like imaginary lines have no authority	1	1	0.99
P_2	this is where the police and prisons need to point out who these ppl are. an eye for an eye should be the approach and these men shud be beaten day after day after day filthy animals shud not be protected by law	1	1	0.99
P_3	you bastard who the hell do you think you are you are supporting a national sovereignty breach and advocating open borders get the hell out of this country	2	2	0.98
P_4	we stand with saudi arabia my ass go be the president of a different country and suckle the teets of authoritarians you fucking asshole i personally like my journalists alive and in one piece	2	2	0.95
P_5	i took a selfie from my window i had one leg in the air and one on the balcony yes it was dangerous but you know what it was worth it now i have the best selfie and i am still alive bingo	3	3	0.99
P_6	once on read light i can speed like anything to get ahead of anyone i just like to be number one on the road and do not care how much faster i have to drive	3	3	0.99
P_7	i hope you die drowning in medical debt then this joke will be hillarious	4	4	0.99
P_8	you haven t known me long enough to know i love pissing people off she crosses her arms	4	4	0.96
P_9	happy birthday to young rebel star prabhas wish you luck to strike gold with sahoo fans are showering prabhas with well wishes a day before his birthday	0	0	0.99
P_{10}	a funny story while canvassing in md i was wearing my beto shirt lady opens door i say my speil she says i would never vote for beto cruz he s a fraud and i don t wish you luck a ya andb um md	0	0	0.91

TABLE 7. Word embeddings.

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

the ability to generate text representation that is invariant to irrelevant factor and highly precise and relevant to the crucial aspect of online antisocial behaviour is indispensable. The traditional machine learning approach of feature extraction, also known as shallow processing, is limited in nature and works only with surface-level features and well-structured documents, however, can some time prove inadequate to handle challenging user-generated data. Thus, more sophisticated techniques are imperative if subtle and often latent aspects are conclusive in the accurate class assignment. This is where deep learning technology and the word embeddings have proven successful in identifying different types of online antisocial behaviours and, has demonstrated promising results in this paper.

V. CONCLUSION

Online antisocial behaviour is a public health threat and a social problem. It is a prevailing pattern of disregard for, and violation of the rights of others. The exhibition of antisocial behaviour might be an entertaining act for a perpetrator, nonetheless, can lead a victim into anxiety, depression, low self-esteem, self-confinement, and suicidal ideation. Twitter and other online platforms can sometimes become incubators for such behaviour, leading to numerous societal problems. Given a large amount of unstructured social media data, a scalable and robust automatic tweet classification technique is imperative for the efficient management of online content, and for the timely intervention by the platforms to prevent dire situations. The paper proposes an approach for multi-class identification of antisocial behaviour from social media posts using the state-of-the-art deep learning models. Its main contributions are:

1)Benchmark medium-scale antisocial behaviour dataset with multi-class annotation; 2) Development of a deep learning classification model succeeding performance evaluation against its different architectures; 3) Performance validation of the deep learning model against traditional machine learning baselines; 4) Error analyses, using visually enhanced

graphical interpretation of similarities among the different classes of antisocial behaviour, to intercept the sources of misclassifications; 5) Knowledge discovery related to antisocial behaviour tweets by leveraging descriptive analyses.

An extensive set of experiments were conducted implementing different feature-model combinations of deep learning architectures and the results are presented in Table-4. Overall, the deep learning models with GloVe embedding achieved higher scores in all evaluation metrics compared to the same models using Word2Vec embeddings and also to the traditional machine learning algorithms (except for RNNs). The highest accuracy was achieved by GRUs using GloVe embeddings, with a batch size of 32, and Nadam optimizer. Along with the empirical validation of the superiority of the proposed deep learning model over the traditional machine learning algorithms, a realistic solution for a real-world problem of antisocial behaviour has been presented. As a part of the study, a benchmark antisocial behaviour corpus was created by manually annotating the tweets under the supervision of domain experts. Such corpora can reduce the time and the cost needed to prepare antisocial behaviour dataset for future studies, and the importance of which has been emphasized in [83]. To better understand the inner working of the classification process and to analyze errors, the dimensionality reduction set of scatter plots were created to demonstrate both performance and misclassification in 2D space. To further quantify the classification scores of each class, the confusion matrices were generated to complement the analyses and to pinpoint the main root cause of misclassifications. The matrices also provided decision support for choosing the optimal model for a specific ASB category detection. For example, GRU performed better when detecting Class-3 and Class-0 tweets, and both LSTM's and GRU's performance was comparable in detecting the remaining three classes

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

REFERENCES

[1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Washington, DC, USA: American Psychiatric Pub, 2013.

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

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

[4] D. L. Haynie, E. Silver, and B. Teasdale, "Neighborhood characteristics, peer networks, and adolescent violence," *J. Quant. Criminol.*, vol. 22, no. 2, pp. 147–169, Jun. 2006, doi: 10.1007/s10940-006-9006-y.

[5] A. M. Gard, H. L. Dotterer, and L. W. Hyde, "Genetic influences on antisocial behavior: Recent advances and future directions," *Current Opinion Psychol.*, vol. 27, pp. 46–55, Jun. 2019.

[6] S. B. Manuck and J. M. McCaffery, "Gene-environment interaction," *Annu. Rev. Psychol.*, vol. 65, no. 1, pp. 41–70, Jan. 2014, doi: 10.1146/annurev-psych-010213-115100.

[7] L. W. Hyde, R. Waller, C. J. Trentacosta, D. S. Shaw, J. M. Neiderhiser, J. M. Ganiban, D. Reiss, and L. D. Leve, "Heritable and nonheritable pathways to early callous-unemotional behaviors," *Amer. J. Psychiatry*, vol. 173, no. 9, pp. 903–910, Sep. 2016.

[8] R. Singh, J. Du, Y. Zhang, H. Wang, Y. Miao, O. A. Sianaki, and A. Ulhaq, "A framework for early detection of antisocial behavior on Twitter using natural language processing," in *Proc. Conf. Complex, Intell., Softw. Intensive Syst.*, 2019, pp. 484–495.

[9] M. S. 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.

[10] E. V. Altay and B. Alatas, "Detection of cyberbullying in social networks using machine learning methods," in *Proc. Int. Congr. Big Data, Deep Learn. Fighting Cyber Terrorism (IBIGDELFT)*, Dec. 2018, pp. 87–91.

[11] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, vol. 2, Dec. 2011, pp. 241–244.

[12] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. D. Jong, "Improving cyberbullying detection with user context," in *Proc. Eur. Conf. Inf. Retr.*, 2013, pp. 693–696.

[13] 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*, 2012, pp. 695–698.

[14] S. Subramani, H. Wang, H. Q. Vu, and G. Li, "Domestic violence crisis identification from Facebook posts based on deep learning," *IEEE Access*, vol. 6, pp. 54075–54085, 2018, doi: 10.1109/access.2018.2871446.

[15] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.

[16] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975, doi: 10.1145/361219.361220.

[17] M. Peng, Q. Xie, H. Wang, Y. Zhang, and G. Tian, "Bayesian sparse topical coding," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 6, pp. 1080–1093, Jun. 2019, doi: 10.1109/tkde.2018.2847707.

[18] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Univ. Texas Austin, Austin, TX, USA, Tech. Rep. 121, 2015.

[19] 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, doi: 10.1613/jair.4272.

[20] 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, 2012, pp. 90–94.

[21] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency," *Freq. Emergence Linguistic Struct.*, vol. 45, p. 137, 2001.

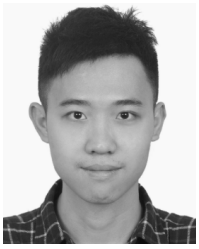
[22] Z.-G. Chen, Z.-H. Zhan, H. Wang, and J. Zhang, "Distributed individuals for multiple peaks: A novel differential evolution for multimodal optimization problems," *IEEE Trans. Evol. Comput.*, vol. 24, no. 4, pp. 708–719, Aug. 2020.

[23] Y. H. Zhang, Y. J. Gong, Y. Gao, H. Wang, and J. Zhang, "Parameter-free Voronoi neighborhood for evolutionary multimodal optimization," *IEEE Trans. Evol. Comput.*, vol. 24, no. 2, pp. 335–349, Apr. 2020.
[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[25] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1017–1024.
[26] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882
[27] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, *arXiv:1404.2188*. [Online]. Available: http://arxiv.org/abs/1404.2188
[28] 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.
[29] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," 2016, *arXiv:1605.05101*. [Online]. Available: https://arxiv.org/abs/1605.05101
[30] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:1308.0850*. [Online]. Available: http://arxiv.org/abs/1308.0850
[31] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.
[32] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*. [Online]. Available: http://arxiv.org/abs/1409.1259
[33] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. Int. Conf. Artif. Neural Netw.*, 2005, pp. 799–804.
[34] 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.
[35] 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 (WWW Companion)*, 2017, pp. 759–760.
[36] 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. 137–147.
[37] 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, vol. 31, no. 3, pp. 632–635.
[38] 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. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 4343–4352.
[39] S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang, and H. Shakeel, "Deep learning for multi-class identification from domestic violence online posts," *IEEE Access*, vol. 7, pp. 46210–46224, 2019, doi: 10.1109/access.2019.2908827.
[40] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," 2017, *arXiv:1702.01923*. [Online]. Available: http://arxiv.org/abs/1702.01923
[41] G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. P. Hubbard, R. J. B. Dobson, and R. Dutta, "Characterisation of mental health conditions in social media using informed deep learning," *Sci. Rep.*, vol. 7, no. 1, p. 45, Apr. 2017.
[42] J. Trofimovich, "Comparison of neural network architectures for sentiment analysis of Russian tweets," in *Proc. Int. Conf. Dialogue*, 2016, pp. 50–59.
[43] 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.
[44] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Proc. Eur. Conf. Inf. Retr.*, 2018, pp. 141–153.
[45] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," 2019, *arXiv:1902.06673*. [Online]. Available: http://arxiv.org/abs/1902.06673
[46] A. Agrawal, "Clickbait detection using deep learning," in *Proc. 2nd Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Oct. 2016, pp. 268–272.
[47] J. Seering, R. Kraut, and L. Dabbish, "Shaping pro and anti-social behavior on twitch through moderation and example-setting," in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2017, pp. 111–125.
[48] P. Liu, J. Guberman, L. Hemphill, and A. Culotta, "Forecasting the presence and intensity of hostility on Instagram using linguistic and social features," in *Proc. 12th Int. AAAI Conf. Web Social Media*, 2018, pp. 181–190.
[49] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proc. 1st Workshop Trolling, Aggression Cyberbullying (TRAC)*, 2018, pp. 1–11.
[50] F. K. Ventirozos, I. Varlamis, and G. Tsatsaronis, "Detecting aggressive behavior in discussion threads using text mining," in *Proc. Int. Conf. Comput. Linguistics Intell. Text Process.*, 2017, pp. 420–431.
[51] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, "Trolls just want to have fun," *Personality Individual Differences*, vol. 67, pp. 97–102, Sep. 2014, doi: 10.1016/j.paid.2014.01.016.
[52] P. Shachaf and N. Hara, "Beyond vandalism: Wikipedia trolls," *J. Inf. Sci.*, vol. 36, no. 3, pp. 357–370, Jun. 2010, doi: 10.1177/0165551510365390.
[53] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2017, pp. 1217–1230.
[54] J. Guberman and L. Hemphill, "Challenges in modifying existing scales for detecting harassment in individual tweets," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 1–10.
[55] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychol. Bull.*, vol. 140, no. 4, pp. 1073–1137, Jul. 2014, doi: 10.1037/a0035618.
[56] B. Cao and W.-Y. Lin, "How do victims react to cyberbullying on social networking sites? The influence of previous cyberbullying victimization experiences," *Comput. Hum. Behav.*, vol. 52, pp. 458–465, Nov. 2015.
[57] S. Herring, K. Job-Sluder, R. Scheckler, and S. Barab, "Searching for safety online: Managing 'trolling' in a feminist forum," *Inf. Soc.*, vol. 18, no. 5, pp. 371–384, Oct. 2002, doi: 10.1080/01972240290108186.
[58] N. Sest and E. March, "Constructing the cyber-troll: Psychopathy, sadism, and empathy," *Personality Individual Differences*, vol. 119, pp. 69–72, Dec. 2017, doi: 10.1016/j.paid.2017.06.038.
[59] C. V. Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. D. Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PloS one*, vol. 13, no. 10, 2018, Art. no. 203794.
[60] N. Tahmasbi and A. Fuchsberger, "Challenges and future directions of automated cyberbullying detection," in *Proc. Amer. Conf. Inf. Syst.* Atlanta, GA, USA: Assocation of Information Systems, 2018, pp. 1–10.
[61] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the Instagram social network," in *Proc. Detection Cyberbullying Incidents Instagram Social Netw.*, 2015, pp. 49–66.
[62] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection with participant-vocabulary consistency," *Social Netw. Anal. Mining*, vol. 8, no. 1, p. 38, Dec. 2018, doi: 10.1007/s13278-018-0517-y.
[63] S. T. Aroyehun and A. Gelbukh, "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling," in *Proc. 1st Workshop Trolling, Aggression Cyberbullying (TRAC)*, 2018, pp. 90–97.
[64] N. Craker and E. March, "The dark side of Facebook: The dark tetrad, negative social potency, and trolling behaviours," *Personality Individual Differences*, vol. 102, pp. 79–84, Nov. 2016, doi: 10.1016/j.paid.2016.06.043.
[65] J. De-La-Pena-Sordo, I. Santos, I. Pastor-López, and P. G. Bringas, "Filtering trolling comments through collective classification," in *Proc. Int. Conf. Netw. Syst. Secur.*, 2013, pp. 707–713.
[66] E. E. Buckels, P. D. Trapnell, T. Andjelovic, and D. L. Paulhus, "Internet trolling and everyday sadism: Parallel effects on pain perception and moral judgment," *J. Personality*, vol. 87, no. 2, pp. 328–340, Apr. 2019, doi: 10.1111/jopy.12393.
[67] E. March, R. Grieve, J. Marrington, and P. K. Jonason, "Trolling on tinder (and other dating apps): Examining the role of the dark tetrad and impulsivity," *Personality Individual Differences*, vol. 110, pp. 139–143, May 2017, doi: 10.1016/j.paid.2017.01.025.
[68] L. Li, T.-T. Goh, and D. Jin, "How textual quality of online reviews affect classification performance: A case of deep learning sentiment analysis," *Neural Comput. Appl.*, vol. 32, no. 9, pp. 4387–4415, May 2020.

[69] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," *Inf. Process. Manage.*, vol. 57, no. 1, 2020, Art. no. 102121, doi: 10.1016/j.ipm.2019.102121.
[70] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning based text classification: A comprehensive review," in *Proc. Deep Learn. Based Text Classification, Comprehensive Rev.*, 2020, pp. 1–37.
[71] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[72] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[73] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*. [Online]. Available: http://arxiv.org/abs/1607.01759
[74] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," 2014, *arXiv:1412.1058*. [Online]. Available: http://arxiv.org/abs/1412.1058
[75] 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.
[76] N. Reimers and I. Gurevych, "Optimal hyperparameters for deep lstm-networks for sequence labeling tasks," 2017, *arXiv:1707.06799*. [Online]. Available: http://arxiv.org/abs/1707.06799
[77] F. Chollet, "Keras: Deep learning library for theano and tensorflow," Tech. Rep., 2015.
[78] M. Abadi, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement.*, 2016, pp. 265–283.
[79] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 1019–1027.
[80] B. Pang, L. Lillian, and V. Shivakumar, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, vol. 10, Jul. 2002, pp. 79–86.
[81] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "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.
[82] 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, doi: 10.1109/72.279181.
[83] 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, doi: 10.1016/j.jbi.2014.11.002.

RAVINDER SINGH received the master's degree (Hons.) in computer science engineering from Santa Clara University, Santa Clara, CA, USA, in 2015. He is currently pursuing the Ph.D. degree with the Institute for Sustainable Industries and Liveable Cities, Victoria University, Footscray, VIC, Australia. His research interests include deep learning, data mining, natural language processing, social media analysis, and behavior studies. He is particularly interested in the application of deep learning in detecting personality disorders from online corpora. He is a member of the Golden Key International Honour Society. He was a recipient of Australian government's Research Training Program Scholarship.

SUDHA SUBRAMANI received the B.E. and M.E. degrees in computer science from Anna University, India, in 2010 and 2012, respectively, and the Ph.D. degree from Victoria University, Australia. Her research affiliation is with the Institute for Sustainable Industries and Liveable Cities, Victoria University, where she is currently working as a Lecturer with the College of Engineering and Science. She has completed a 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 a Gold Medal for her academic excellence in the master's degree, and was also a recipient of the International Postgraduate Research Scholarship.



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 with the Institute for Sustainable Industries and Liveable Cities, Victoria University. His research interests include machine learning, deep learning, data mining, natural language processing, social media analysis, and their applications. He is particularly interested in quality evaluation, knowledge discovery, and content recommendation on online user-generated reviews.

YANCHUN ZHANG (Member, IEEE) has been a Professor and the Director of the Centre for Applied Informatics, Victoria University, since 2004. He has published more than 300 research papers in international journals and conference proceedings, including *ACM Transactions on Computer–Human Interaction (TOCHI)*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), VLDBJ, SIGMOD, and ICDE conferences, as well as health/medicine journals. His research interests include databases, data mining, Web services, and e-health. He is currently a Founding Editor and the Editor-in-Chief of *World Wide Web Journal* (Springer) and *Health Information Science and Systems* Journal (Springer). He has served as expert panel member at various international funding agencies, such as Australia Research Council, National Natural Science Fund of China (NSFC), the Royal Society of New Zealand's Marsden Fund from 2015 to 2017, Performance-Based Research Fund (PBRF) Panel of New Zealand, Medical Research Council of U.K., and NHMRC of Australia on Built Environment and Prevention Research.

HUA WANG (Member, IEEE) 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 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. He is currently a full time Professor with Victoria University. As a Chief Investigator, three Australian Research Council (ARC) Discovery grants have been awarded since 2006, and 280 peer-reviewed scholar articles have been published. Ten Ph.D. students have already graduated under his principal supervision.

KHANDAKAR AHMED (Member, IEEE) received the M.Sc. degree in networking and ebusiness centered computing (NeBCC) under the joint consortia of University of Reading, U.K., the Aristotle University of Thessaloniki, Greece, and the Charles III University of Madrid (UC3M), Spain. He has a decade of the comprehensive academic experience of working across South Asia, Europe, and Australasia. He has extensive industry engagement as a Chief Investigator in multiple research projects related to the Internet-of-Things, smart cities, machine learning, cybersecurity, and biomedical informatics. He is currently a Lecturer with the Discipline of IT, College of Engineering and Science, Victoria University. He has published more than 50 journal articles and conference proceedings, including three book chapters. His current research interests include the application of machine learning across biomedical informatics, the Internet of Things, smart technology, and cybersecurity.

ZHENXIANG CHEN (Member, IEEE) received the B.S. and M.S. degrees from the University of Jinan, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree from the School of Computer Science and Technology, Shandong University, Jinan, in 2008. He is currently a Professor with the School of Information Science and Engineering, University of Jinan. His research interests include network behavior analysis, mobile security and privacy, and hybrid computational intelligence. He has authored numerous articles in these areas.