Detection of Hate Tweets using Machine Learning and Deep Learning

Lida Ketsbaia, Biju Issac and Xiaomin Chen
Computer and Information Sciences
Northumbria University
Newcastle-upon-Tyne, UK
{lida.ketsbaia, biju.issac, xaiomin.chen}@northumbria.ac.uk

***Abstract*— Cyberbullying has become a highly problematic occurrence due to its potential of anonymity and its ease for others to join in the harassment of victims. The distancing effect that technological devices have, has led to cyberbullies say and do harsher things compared to what is typical in a traditional face-to-face bullying situation. Given the great importance of the problem, detection is becoming a key area of cyberbullying research. Therefore, it is highly necessary for a framework to accurately detect new cyberbullying instances automatically. To review the machine learning and deep learning approaches, two datasets were used. The first dataset was provided by the University of Maryland consisting of over 30,000 tweets, whereas the second dataset was based on the article 'Automated Hate Speech Detection and the Problem of Offensive Language' by Davidson et al., containing roughly 25,000 tweets. The paper explores machine learning approaches using word embeddings such as DBOW (Distributed Bag of Words) and DMM (Distributed Memory Mean) and the performance of Word2vec Convolutional Neural Networks (CNNs) to classify online hate.**

Keywords — Hate Speech; CNN, Machine Learning; Word2Vec; Doc2Vec

I. INTRODUCTION

Web 2.0 plays a distinct role within relationships and communication in today's society. While most individuals use the Internet as a harmless and beneficial method of interaction and communication, some identify it as a method of anonymity and freedom to express themselves without the fear of face to face interaction. This has led to bullying proliferating as technology evolved. There are online datasets of hate speech available for research [1] and [2].

The development of social media has steered people to adopting a new method of spreading hate. Bullying represents a type of aggression that takes on various forms, such as physical, verbal, and relational. During the mid-2000's a new genre of peer aggression was identified called cyberbullying, which took place using digital or online means. Smith et al, [3] defined cyberbullying as 'an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself". Due to the recent popularity and growth of social media platforms such as Twitter, cyberbullying is becoming more and more prevalent.

In contrast to traditional bullying, cyberbullying is not limited to a place and time. A concern many researchers have is that victims do not perceive their experiences as bullying, leading to many victims not reporting such instances or seek the required help for their emotional destress. Numerous studies have supported these statements, [4] reported that approximately 90% of young cyberbullying victims did not tell their parents or other trusted adults about their negative online experiences. Such factors are increasingly worrying as many victims often deal with both psychiatric and psychosomatic disorders [5] and in worst case suicide [5]. A British study found that nearly half of suicides among young people were related to bullying [6]. Additionally, Google searches of bullying have increased threefold since 2004 [7]. These facts identify an urgent need to apprehend, identify and reduce its widespread presence.

One of the most popular social mediums found today is Twitter, a micro-blogging site that enables users to write up to 280 characters of text commonly referred to as tweets [8]. Advances in Twitter has changed the way people share their feelings and views with a wider audience due to its free format messages and easy accessibility.

Twitter is a real time information networking site that enables the collection of global opinions that is of public interest, allowing Twitter to become an excellent channel to analyse peoples' social interactions and opinions. Cyberbullying through Twitter has received attention in recent years because of its association with a number of tragic, high-profile suicides. [9].

Traditional mechanisms have been implemented to tackle the issue of cyberbullying within Social Media platforms, with companies incorporating guidelines that their users must follow, as well as employing editors to manually check for bullying behaviour. However, these methods have fallen short in tackling the issue since maintaining such mechanisms is both time and labour consuming. A study [10] on college students found that 69.4% were actively using the micro-blogging website Twitter, with 45.5% reporting cyberbullying. Within their study the prevalence of cyberbullying on Twitter was higher than other social media platforms including but not limited to Facebook (38.6%), Instagram (13.7%), and YouTube (11.4%).

In this work different machine learning and deep learning techniques are applied as a comparison method to detect online harassment within the social media platform Twitter.

978-1-6654-0392-4/20/$31.00 ©2020 IEEE
DOI 10.1109/TrustCom50675.2020.00103

II. Related work

A. Machine Learning

One of the first studies that examined the effectiveness of machine learning in respect to sentiment analysis was Pang and Vaithyanathan [11]. The researchers wanted to classify movie reviews by sentiment through a negative and positive scale. Their results identified that Naïve Bayes had the worst performance whereas SVMs had the best although the differences between the two were not extensive. Moreover, [11] found one common phenomenon within reviews which caused Machine Learning approaches to misinterpret sentences and identified it as "thwarted expectations [11]. Thwarted expectations are narratives that an author would deliberately create causing a contrast between his words and his thoughts.

Twitter specific analysis has occurred throughout the years, however there is a difference between a normal sentiment analysis and a Twitter sentiment analysis, this is due to the shortness of twitter posts. Twitter messages tend to use slang words and misspellings since a maximum of 280 characters is allowed. Zhao and Mao [13] reported the use of an embedding-enhanced bag-of-words approach to detect cyberbullying through participant-vocabulary consistency. Other efforts have focused on the use of complementary information to enhance text-based cyberbullying detection [14]. Menger, Scheepers and Spruit [15] presented an improved model using user-based features, i.e., the history of the user's activities and demographic features. Huang, Singh and Atrey focused on social network features for cyberbullying content and provided improved performance in cyberbullying detection by considering online relationships [16].

SVM has been progressively used to develop bullying prediction models, researchers found that incorporating SVM was increasingly effective. For example, Chen et al. [17] used a Social Media dataset that included cyberbullying instances, furthermore a SVM cyberbullying prediction model was applied to detect whether the content was offensive. Their results indicated that SVM is more accurate in detecting offensiveness in comparison to Naïve Bayes (NB), however, their Naïve Bayes predictions were faster than SVM. Chavan and Shylaja [18] proposed a similar framework using a dataset containing offensive words, SVM was used to build a classifier to detect cyberbullying within their dataset. The results concluded that the SVM classifier detected cyberbullying more accurately than Logistic Regression. A paper created by Mangaonkar et al. [19] collected data from YouTube, their results suggested that the SVM cyberbullying model is more reliable but not as accurate as rule based Jrip. However, the SVM-based cyberbullying model was more accurate than NB.

Furthermore, studies have also focused on cyberbullying prediction based on irreverent words as a feature set [20], [21], [22], [23], [24]. Lexicons containing profane words have been created to indicate bullying and have been used as features for input to machine learning algorithms [25], [26]. Research has shown that using profane words as features demonstrates a significant improvement within machine learning model performance.

B. Deep Learning

Although neural networks have existed for several years, it was not until the last decade that they have been used competitively in dealing with real word problems. Due to its processing capabilities and the advent of fast graphics processing units. The main arguments for the use of deep learning is its proficiency to automatically identify and extract features, therefore achieving a higher accuracy and performance. In general, the hyperparameters of classifier models are also measured automatically. In contrast to deep learning, machine learning features are defined and extracted either manually or by using feature selection methods.

In recent years, there have been two deep learning architectures that have frequently outperformed:

Recurrent Neural Networks (RNN): RNN is described as "class of neural networks whose connections between neurons form a directed cycle, which creates feedback loops within the RNN. The main function of RNN is the processing of sequential information on the basis of the internal memory captured by the directed cycles. Unlike traditional neural networks, RNN can remember the previous computation of information and can reuse it by applying it to the next element in the sequence of inputs" [27]. Convolutional Neural Networks (CNN): Typically employed within areas such as computer vision and NLP. CNN is identified as a feed-forward neural network, its architecture is composed of convolutional and pooling or subsampling layers. These layers would then provide inputs to a fully connected classification layer [27]. Dang, Moreno-García and De la Prieta expressed that "Convolution layers filter their inputs to extract features; the outputs of multiple filters can be combined. Pooling or subsampling layers reduce the resolution of features, which can increase the CNN's robustness to noise and distortion. Fully connected layers perform classification tasks"[27].

A study [28] discussed how Deep Learning techniques, such as Word Embeddings in addition to Recurrent Neural Networks, have shown to have a greater potential than typical machine learning methods. Within their work they applied different Deep Learning and machine learning techniques to predict violent incidents during psychiatric admission using clinical text. Their results identified that using Deep Learning provided an improved performance in comparison to Machine learning. Ali, El Hammid and Yousif's research introduced a developed classification sentiment analysis using deep learning networks and compared the results of different deep learning networks. The researchers found that deep learning greatly outperformed Naïve Bayes and SVM. [29]

C. Current Issue faced in detecting Cyberbullying

Large volumes of work regarding sentiment analysis has been conducted throughout the past two decades and continues to rapidly grow in various directions, newer research depends on developing more accurate sentiment classifiers using machine learning however challenges on the work still remain. For

example: Negation is of high importance since one negation word could largely impact the polarity of a sentence making it from positive to negative or vice versa. Sarcasm plays a problematic role when assigning a label of sentiment as it can wrongly indicate a user's emotional state. Quotes and retweets are largely difficult to assess since it is often unclear and not explicitly evident as to whether the person that retweeted or quoted a specific sentence has the same stance as the person who was retweeted/quoted. A person's emotional state may or may not have the same polarization as the opinion that he or she tries to express. A user may report information without an indication of the emotional state they are going through, causing an unclear consideration for the statement causing a positive or negative identification to be produced inaccurately. Due to the shortness of Twitter messages, the occurrence of incorrect spelling and the use of slang words is more often than in any other domains, therefore such use of acronyms, misspelled words and emoticons can cause an issue when trying to classify the sentiment state of a message

III. Datasets

A dataset created by the University of Maryland was used, affiliated with the paper "A Large Human-Labeled Corpus for Online Harassment Research". The purpose of the data was to "help train machine learning models, identify the linguistic features of online harassment and for studying the nature of harassing comment and the culture of trolling" [2].

TABLE 1. MARYLAND UNIVERSITY DATASET EXAMPLES

Hate	**Non-Hate**
lmfaaaooooo you fucking nigger	The Jews of South Africa, who are mostly Ashkenazi Jews, descended from pre-Holocaust immigrant Lithuanian Jews.
Untag you fucking nigger	Actually, the Nazis just wanted rid of the Jews. Didn't really care where they went.
An anti-white will deny that 2 2=4 and the sky is blue if it suits them. #WhiteGenocide	Can anyone wear the Star of David, that the Jews have reclaimed? No, so why do white people disrespect what "nigga" means
That's fine, as long as those immigrants are White. #WhiteGenocide	Thanks be to the Jews of Israel
YES He did. Back in the glory days of the KKK and ALL the closet democrat elected in DC. #EVIL, #WhitePower	Thank you @SenTedCruz "If you will not stand with Israel, you will not stand with the Jews, I will not stand with… https: /

The dataset deals with violent online harassment, that contains but is not limited to the use of violent/sexually violent phrases, threats as well as racist, hateful, and derogatory comments. A list of search terms was used to download relevant data from Twitters API which included terms as follows (for example): #whitgennocide, #fuckniggers, #WhitePower, #WhiteLivesMatter, Feminist and The Jews.

The authors ensured that the tweets collected were amongst the worst, identifying the most offensive or violent messages which included largely racist/misogynistic and homophobic tweets, overall, they were messages that could be upsetting to the general reader. Such depth enables users to understand and evaluate the true extent of hate that can be seen online. The labeled corpus took three months to manually label, creating a dataset of 35,000 tweets. Table 1 provides an example of the "Hate" Tweets and "Non-Hate" Tweets the dataset consisted of.

The second dataset used was based on the article "Automated Hate Speech Detection and the Problem of Offensive Language" [2]. Using Twitter's API the authors searched for tweets containing terms from a lexicon resulting in tweets from 33,458 Twitter users. A random sample of 25,000 tweets were chosen and Crowd Flower workers were asked to individually label the tweets as one of three categories: hate speech, offensive (not hate) or neither offensive nor hate speech. Based on the majority decision for each tweet a label was assigned. In conclusion the resulting sample consisted of 24,802 labeled tweets. Only 5% of tweets were coded as hate speech 76% were identified as offensive language and the remainder were non-offensive. Due to the research specifically targeting hate speech this dataset was altered to only include the "hate speech" and "non-offensive or non-hate" tweets as in table 2.

TABLE 2. CONRELL UNIVERSITY DATASET EXAMPLES

Hate	**Non-Hate**
"hs a beaner smh you can tell hes a mexican	birds outside my bedroom window have way too much to talk about for first thing in the morning!
you're fucking gay, blacklisted hoe" Holding out for #TehGodClan anyway https://t.co/xUCcwoetmn	I don't think I even been in a real relationship…i thought they was real but they were just trash
LMFAOOOO I HATE BLACK PEOPLE This is why there's black people and niggers	World Cup 2014 diary: Argentina mock Brazil with twirling towels;
At least I'm not a nigger https://t.co/RGJa7CfoiT""	Confused. Are lefties using the #Bridgeghazi hashtag to mock the bridge situation as frivolous, or seriously equating the t…
#Dutch people who live outside of #NewYorkCity are all white trash.	Me and … are in our yellow submarine rn.. 🌊

The following research is performed as a starting base of what works with the datasets provided and how it can incorporate certain deep learning and machine learning models.

IV. Pre-processing

Given that the research focuses on Twitter messages the requirement of pre-processing is largely necessary as tweets tend to not be formatted in the required way needed for a text analysis to occur. Kappas et al., [30] distinguished that "Text based communication in English seems to frequently ignore the rules of grammar and spelling" therefore making it necessary for preprocessing to be required to produce a cleaner dataset thus increasing the performance of classification that are later used significantly.

The following preprocessing techniques were used prior to the classification analysis:

- Removal of numbers: numbers are removed from the dataset as they do not carry any sentiment, however, there are some researchers that argue this method.
- Lowercasing: lowercasing is one of the more common pre-processing techniques and sometimes overlooked. By doing words are merged and the dimensionality of the problem is reduced
- Replacing URLs and user mentions: Tweets that include various links or URLS do not contribute to the sentiment of a tweet, therefore they were parsed and replaced, additionally usernames which are used to refer too other users with the @ symbol are again not required therefore removed during the preprocessing stage for the experiment.
- Replacing Contractions: for example: "can't" will be replaced as cannot.
- Punctuation and special characters removal: punctuations and symbols were removed even though there are instances in which a punctuation mark can denote an existence of a sentiment either negative or positive.
- Decoding HTML (i.e '&','&quot', etc.).

A. *Imbalanced Datasets*

To deal with the problem of imbalanced datasets Sklearn resample was implemented to resample the minority and majority labels to either fit the minority or majority class, thus presenting a balanced dataset.

B. *Removal of Stopwords*

When a document is collected each individual term found within the document plays a substantial role in understanding whether it fits to a specific category, thus it is commonly practiced to remove common functional terms. These terms are identified as "stopwords" and take the form of "the", "but" "if" etc. Regardless of having a grammatical function they do not reveal anything regarding the content found within the documents. Researchers have commonly removed "stopwords" as a hope to increase retrieval. Most textual data found in documents are designed to have a syntactic role rather than a semantic one therefore by removing "stopwords", enables a more thorough understanding of the text and its sentiment. However, over the years "stopwords" have recently been an area of debate of whether they might hold a substantial value when categorizing data.

V. MACHINE LEARNING CLASSIFICATION

A. *Classifiers*

To get a better understanding of how different classifiers work, four different classifiers where applied whilst incorporating various ngram ranges: unigrams, bigrams, and trigrams:

Logistic Regression: Allows a way to combine pieces of contextual evidence to estimate the probability of a certain class occurring within a certain context. Its task is to estimate the probability of class 'y' occurring within context 'X'.

Linear SVC: When in an ideal situation classes would be linearly separable, since the feature space would be divided into class segments by creating a hyperplane finding the largest margin of the two classes in the training set. The closest data points for both classes found parallel to the hyperplane would constitute as the support vectors. Therefore, SVM attempts find the best possible surface to separate positive and negative training samples [32]. SVM has largely been used to build cyberbullying prediction models and have so far found to be effective and efficient [33],[17].

Naive Bayes: NB has increasingly been implemented to construct cyberbullying prediction and can be found in models produced by numerous researchers [34],[35],[36],[37]. "This model assumes that the text is generated by a parametric model and utilizes training data to compute Bayes-optimal estimates of the model parameters". The paper focuses on two Naïve Bayes models: Multinomial Naïve Bayes: The purpose of th model is to determine the number of times a term occurs within a document (term frequency). Since a term plays a substantial role when deciding the sentiment of a given document, Multinomial Naïve Bayes would be a good choice within classification. Term frequency is helpful whilst deciding if a term would be useful within the analysis or not [38]. Bernoulli Naïve Bayes: Features are independent binary variables as it will indicate the presence or absence of a feature (1 and 0). The difference between Multinomial and Bernoulli is that the multinomial approach takes into consideration the term frequencies whereas Bernoulli approach is interested in concocting whether a term is present or absent in the document under consideration [38].

B. *Metrics*

Upon the classifiers being constructed for unigrams, bigrams and trigrams, an evaluation of the models is produced. The reason behind this is to get a deeper understanding of the classifier's behavior over a global accuracy that will mask the weaknesses within one class of a multiclass problem. The classification report is used to compare all the classification models used throughout the report and choose the ones that have a stronger classification metrics or the ones that are more balanced. The metrics will be defined in the terms of true and false positives as well as true and false negatives. A true positive is when the actual class is positive as well as the estimated class, whereas as a false positive shows an actual class of negative but an estimated class of positive. The performance evaluation is as follows [39]:

Accuracy: Identifies the total number of predictions that were correct

$$Precision = \frac{\text{True Positive (TP)} + \text{True Negative(TN)}}{\text{Total Number of Observations}}$$

Precision: Denotes the proportion of predicted positive cases that are correctly true Positives. The ability of a classifier to not label a positive when in fact is negative.

$$Precision = \frac{true\ positive}{true\ positive + fasle\ positive}$$

Recall: The proportion of actual positive being identified correctly is given by recall

$$Recall = \frac{true\ postive}{true\ positive + false\ negative}$$

F1 score: When reading literature on Precision and Recall, F1 score cannot be avoided. F1 is needed when trying to seek a balance between the Precision and Recall and when there is an uneven class distribution. Therefore the F1score is a weighted mean of the two factors in with the best score being 1.0 and the worst 0.0.

$$F1 = 2\ x\ \frac{precision * recall}{precision + recall}$$

The results identified that throughout both datasets LinearSVC had the highest accuracy whilst incorporating all three different n-gram ranges. Table 3 and Table 4 show the accuracy of the four different classifiers used across unigrams, bigrams and trigrams for Dataset 1 and Dataset 2, respectively. However, Dataset 2 produced higher results ranging between 80.72% - 91.57% whereas Dataset 1 ranged between 67%-80.33

TABLE 3: DATASET 1 - ACCURACY, PRECISION, RECALL AND F1 SCORE OF THE FOUR DIFFERENT CLASSIFIERS USED ACROSS UNIGRAMS, BIGRAMS AND TRIGRAMS

Dataset 1				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
Unigrams				
Logistic Regression	73.00%	72.83%	72.98%	72.87%
Linear SVC	80.00%	80.00%	79.51%	79.67%
Multinomial	75.00%	74.87%	74.50%	74.61%
Bernoulli	74.33%	74.12%	74.20%	74.15%
Bigrams				
Logistic Regression	73.67%	73.58%	73.78%	73.58%
Linear SVC	80.33%	80.80%	79.56%	79.84%
Multinomial	70.00%	72.63%	71.37%	69.81%
Bernoulli	69.33%	74.85%	71.32%	68.66%
Trigrams				
Logistic Regression	74.00%	73.85%	74.02%	73.88%
Linear SVC	80.33 %	80.57%	79.69%	79.92%
Multinomial	69.67%	71.95%	70.94%	69.52%
Bernoulli	67.00%	73.42%	69.19%	66.02%

TABLE 4: DATASET 2 - ACCURACY, PRECISION, RECALL AND F1 SCORE OF THE FOUR DIFFERENT CLASSIFIERS USED ACROSS UNIGRAMS, BIGRAMS AND TRIGRAMS

Dataset 2				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
Unigrams				
Logistic Regression	89.16%	90.24%	89.26%	89.10%
Linear SVC	90.36%	91.84%	90.48%	90.29%
Multinomial	84.34%	86.85%	84.49%	84.11%
Bernoulli	81.93%	85.31%	82.11%	81.54%
Bigrams				
Logistic Regression	90.36%	91.16%	90.45%	90.33%
Linear SVC	91.57%	92.71%	91.67%	91.52%
Multinomial	90.36%	91.84%	90.48%	90.29%
Bernoulli	86.75%	88.63%	86.61%	86.55%
Trigrams				
Logistic Regression	90.36%	91.16%	90.45%	90.33%
Linear SVC	91.57%	92.71%	91.67%	91.52%
Multinomial	89.16%	89.34%	89.20%	89.15%
Bernoulli	80.72%	86.21%	80.49%	79.88%

VI. VECTOR SPACE FOR MEASURING CONTENT

Doc2Vec is a relatively new approach for which NLP uses to obtain vectors. The two main training methods used to obtain Doc2Vec representations are Distributed Memory (DM) and Distributed Bag of Words (DBOW). Distributed Memory can then be further separated Distributed Memory Concatenated (DMC) and Distributed Memory Mean (DMM), the main difference between the two is that DMC concatenates context vectors, whereas the DMM averages them. DBOW forces the model to predict groups of words randomly sampled from the given vector. In practice, DBOW and DM models can be combined to provide other types of vectors.

The research will explore the following:

1. DBOW: Distributed Bag of Words
2. DMM: Distributed Memory model in taking the Mean of context vectors
3. DBOW: Distributed Bag of Words model
4. DBOW+DMM: combination of DBOW and DMM

Using these techniques identified a rise in accuracy, as in table 5 especially for Dataset 1 since the results ranged between 91%-95.33% .DBOW + DMM presented the highest results in both datasets, with Maryland University achieving 95.33% accuracy while using bigrams and Dataset 2 producing 96.39% with trigrams. However, DMM by itself presented the lowest results, whereas bigrams presented the most.

TABLE 5: DOC2VEC RESULTS FOR BOTH DATASETS

Dataset 1			
	Unigrams	**Bigrams**	**Trigrams**
DBOW	93.67%	94.67%	95.33%
DMM	94.66%	94.33%	91.00%
DBOW +DMM	95.00%	95.33%	95.00%
Dataset 2			
	Unigrams	**Bigrams**	**Trigrams**
DBOW	90.37%	91.57%	93.98%
DMM	89.16%	91.57%	86.47%
DBOW + DMM	90.36%	93.98%	96.39%

VII. NEURAL NETWORKS USING DOC2VEC.

Based on the findings produced using Doc2Vec a combination of trigram (DBOW) + bigram (DMM) vectors were created and incorporated within a Neural Network. Various neural networks were created for both datasets to see which would deliver the best results when incorporating

Doc2Vec vectors. Various models were created having between 1-3 hidden layers with either, 64, 128, 256 or 512 hidden nodes. The results identified an accuracy of 96.67% for Dataset 1 and an accuracy of 97.59% for Dataset 2 as in table 6.

VIII. CNN + Word2Vec Model

A. *Word2Vec*

Word2Vec is identified as a Distributed representation of words in a vector space to help learning algorithms achieve a higher performance in NLP tasks by grouping similar words. The Word2Vec model learns word representations is through a pair of architectures: The Continuous Bag-of-Words (CBOW) and Skip-gram.

The CBOW model averages the vectors of all the words within a given context. CBOW is trained is by predicting the current word based upon the projected average of the surrounding context. Skip-gram however predicts surrounding words based on the current word. Words which are a certain distance before and after the current word are predicted with the network being optimized for these predictions [22].

TABLE 6: NEURAL NETWORK TRAINING

		Dataset 1	Dataset 2
Model	**Hidden layer (nodes)**	**Best validation accuracy**	**Best validation accuracy**
Model 1	1 (64)	95.67%	96.39%
Model 2	2 (64)	94.67%	97.59%
Model 3	3 (64)	96.00%	97.59%
Model 4	1 (128)	95.33%	96.39%
Model 5	2 (128)	95.33%	96.39%
Model 6	3 (128)	95.67%	93.98%
Model 7	1 (256)	96.00%	95.18%
Model 8	2 (256)	95.67%	96.29%
Model 9	3 (256)	96.33 %	96.39%
Model 10	1 (512)	96.67%	96.39%
Model 11	2 (512)	96.60%	97.59%
Model 12	3 (512)	96.67%	97.59%
Model 10	1 (512)	96.67%	96.39%

B. *Convolutional Neural Network using Word2Vec*

The two CNN models were created using CBOW and Skip-gram, the architecture of the model was as follows:

- Input layer that defines the length of the input sequences.
- Embedding layer: This layer passes a pre-defined embedding matrix; however, it was made trainable so that it can update the values of vectors as the model trains.
- Three one-dimensional convolutional layers with a kernel size set to 2/3/4 which generate (variable-length) feature maps.
- After each convolutional layer and max pooling layer, it simply concatenates max pooled result from each of the kernel sizes
- Incorporated one fully connected hidden layer with dropout just before the output layer
- Output layer will have just one output node with Sigmoid activation

As per the CNN architecture in figure 1, the results in table 7 detected an increase in accuracy in comparison to the machine learning algorithms throughout both datasets. Based on the CNN architecture using Word2Vec the highest results Dataset 1 produced was through CBOW which witnessed a validation accuracy of 88% and a test accuracy of 89.70% whereas Dataset 2 produced more effective results while using Skip-gram as it had a validation accuracy of 92.78 % and a test of 92.88%.

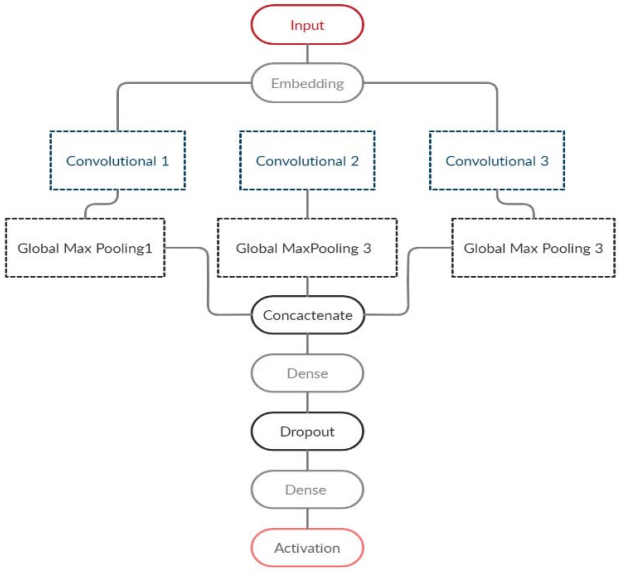

Fig 1. CNN ARCHITECTURE

TABEL 7: CNN USING WORD2VEC AND ROC AUC

Model	Validation	Test	ROC AUC
Dataset 1			
CNN+CBOW	88.00%	89.70%	94.00%
CNN+Skip gram	86.00%	88.04%	92.00%
Dataset 2			
CNN+CBOW	92.77%	89.29%	96.70%
CNN+Skip gram	92.78%	92.88%	98.20%

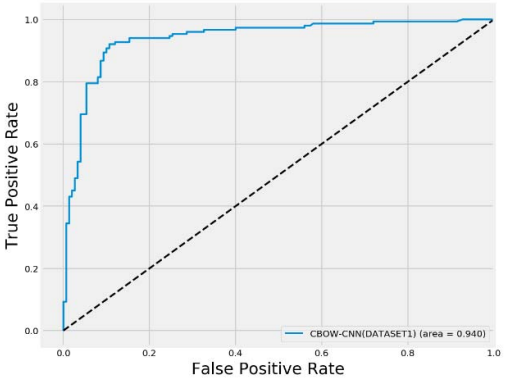

Fig 2: Best ROC AUC for Dataset 1

An AUC-ROC curve was implemented since it is recognized as a performance measurement within classification. ROC is described as a probability curve for while AUC represents measure of separability. It reveals how much a certain model would be able to distinguish between classes. The ROC curve identified a 94% accuracy using CBOW and 92% using Skip-gram for Dataset 1 whereas Dataset 2 produced a 98.60% for Skip-gram and 96.70% of CBOW. The best ROC AUC for dataset 1 and 2 are shown in figures 2 and 3.

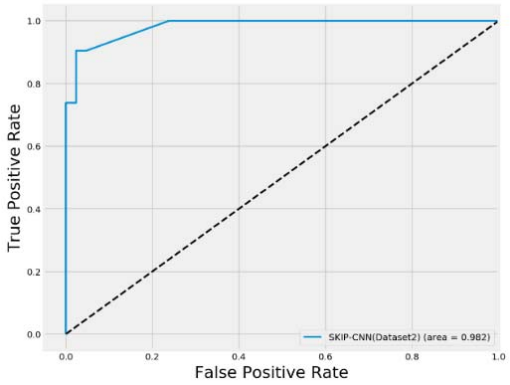

Fig 3: Best ROC AUC for Dataset 2

IX. Comparison with other works

The comparison Table 9 shows how the work conducted compares with others work. The last four rows are our work with different techniques. It clearly shows that the methods adopted work well providing a high accuracy results, however due to the varied approaches on different datasets may not result in a consistent comparison exercise.

Tabel 9: Comparison table

Paper	Technique used	Dataset	Accuracy or F1 Score
Reynolds et al. [40]	J48 and IBK	Kaggle dataset (NUM and NORM)	78.5% and 78.5%
Hani et al. [41]	Neural Network	Kaggle dataset [24] by Reynolds et al.	92.8%
Di Capua et al. [42]	GHSOM	Twitter dataset	72%
Van Hee et al. [43]	Linear SVM	Cyberbullying corpus (English and Dutch)	64% and 61%
X. Zhang et al. [44]	PCNN	Twitter and Formspring.me dataset	98.9% and 96.8%
Our work (Ketsbaia et al.)	Linear SVC	Dataset 1 [1] and Dataset 2 [2]	80.33% and 91.57%
Our work (Ketsbaia et al.)	DBOW + DMM	Dataset 1 [1] and Dataset 2 [2]	95.33% and 96.39%
Our work (Ketsbaia et al.)	CNN + CBOW	Dataset 1 [1] and Dataset 2 [2]	88% and 92.77%
Our work (Ketsbaia et al.)	CNN + Skipgram	Dataset 1 [1] and Dataset 2 [2]	86% and 92.78%
Our work (Ketsbaia et al.)	Neural Network (3 hidden layers) + Doc2Vec	Dataset 1 [1] and Dataset 2 [2]	96.67% and 97.5%

X. Conclusion

During this research, various experiments occurred for our two datasets. Firstly, four different Machine learning algorithms (Logistic Regression, Linear SVC, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes) were used to identify which classification works best with the corpus provided. Throughout both datasets Linear SVC produced the highest results whereas Bernoulli Naïve Bayes produced the lowest. Additionally, it must be acknowledged that Dataset 2 yielded much higher accuracies in comparison to Dataset 1.

Furthermore, the paper proceeded to look at various Doc2Vec models. The models used were DBOW, DMM and a combination of DBOW and DMM. DBOW + DMM presented the best results in both datasets whereas DMM presented the lowest. Interestingly, the accuracy for the dataset created by Maryland University increased drastically and thus had near identical accuracies as Dataset 2. Based upon the results that were produced using Doc2Vec a combination of trigram (DBOW) + bigram (DMM) vectors were utilized to test on a simple Neural Network. Whilst evaluating the neural network an accuracy of 95.33% was achieved for Dataset 1 and an accuracy of 96.38% for Dataset 2.

Lastly two CNN models were developed, the one incorporated CBOW whereas the other implemented Skip-gram. The results generated, were once again drastically higher than the machine learning accuracies, increasing the accuracy by 8% for Dataset 1 and 2% for Dataset 2. ROC AUC was finally employed and identified a 94.00% accuracy for the Maryland University dataset when using CNN + CBOW and a 98.20% for the Cornell University dataset when integrating CNN + Skip-gram. For future work, we will investigate incorporating different Deep Learning models such as Recurrent Neural Networks. Moreover bio-inspired optimization techniques such as Particle Swarm Optimization could be implemented to our current models to see whether it can optimize any of the results.

References

[1] Golbeck, J. et al. “A Large Labeled Corpus for Online Harassment Research”. In Proceedings of the 2017 ACM on Web Science Conference (WebSci '17). ACM, New York, NY, USA, 229-233

[2] T. Davidson, D. Warmsley, M. Michael and I. Weber, Ingmar “Automated Hate Speech Detection and the Problem of Offensive Language” Proceedings of the 11th International AAAI Conference on Web and Social Media, pp.512-515

[3] P. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils", Journal of Child Psychology and Psychiatry, vol. 49, no. 4, pp. 376-385, 2008. Available: 10.1111/j.1469-7610.2007.01846.x.

[4] School Students", Journal of Adolescent Health, vol. 41, no. 6, pp. S22-S30, 2007. Available: 10.1016/j.jadohealth.2007.08.017

[5] L. Beckman, C. Hagquist and L. Hellström, "Does the association with psychosomatic health problems differ between cyberbullying and

traditional bullying?", Emotional and Behavioural Difficulties, vol. 17, no. 3-4, pp. 421-434, 2012. Available: 10.1080/13632752.2012.704228 [Accessed 15 August 2020].
[6] "Fifth of young people bullied in past year - study", BBC News, 2020. [Online]. Available: https://www.bbc.co.uk/news/uk-50370667.
[7] "Cyberbullying Statistics and Facts for 2020 | Comparitech", Comparitech, 2020.
[8] Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data", ACL Anthology, pp. 30-38, 2011.
[9] P. Lee, "Evading the Schoolhouse Gate: Public Schools (K-12) and the Regulation of Cyberbullying", Utah Law Review, vol. 2016, no. 5, pp. 1-60, 2016. [Accessed 26 October 2020].
[10] Calvin, A. Bellmore, J. Xu and X. Zhu, "#bully: Uses of Hashtags in Posts About Bullying on Twitter", Journal of School Violence, vol. 14, no. 1, pp. 133-153, 2014. Available: 10.1080/15388220.2014.966828
[11] Pang, L. Lee and S. Vaithyanathan, "Thumbs up?", Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02, 2002. Available: 10.3115/1118693.1118704 [Accessed 1 June 2020].
[12] Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation", 2018. [Accessed 15 August 2020].
[13] R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation", IEEE Transactions on Fuzzy Systems, vol. 26, no. 2, pp. 794-804, 2018. Available: 10.1109/tfuzz.2017.2690222 [Accessed 15 August 2020].
[14] Raisi and B. Huang. 2016. “Cyberbullying identification using participant-vocabulary consistency” arXiv Preprint arXiv:1606.08084 (2016).
[15] V. Menger, F. Scheepers and M. Spruit, "Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text", Applied Sciences, vol. 8, no. 6, p. 981, 2018. Available: 10.3390/app8060981 [Accessed 15 August 2020].
[16] Q. Huang, V. Singh and P. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis", Proceedings of the 3rd International Workshop on Socially-Aware Multimedia - SAM '14, 2014. Available: 10.1145/2661126.2661133
[17] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, ‘‘Detecting offensive language in social media to protect adolescent online safety,’’ in Proc. Int. Conf. Privacy, Secur., Risk Trust (PASSAT), Sep. 2012, pp. 71–80.
[18] V. S. Chavan and S. S. Shylaja, ‘‘Machine learning approach for detection of cyber-aggressive comments by peers on social media network,’’ in Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI), Aug. 2015, pp. 2354–2358.
[19] Mangaonkar, A. Hayrapetian, and R. Raje, ‘‘Collaborative detection of cyberbullying behavior in Twitter data,’’ in Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT), Dekalb, IL, USA, May 2015, pp. 611–616. [74] H. Sanchez and S. Kumar, ‘‘Twitter bullying detection,’’ Tech. Rep. UCSC ISM245, 2011.
[20] Kontostathis, K. Reynolds, A. Garron, and L. Edwards, ‘‘Detecting cyberbullying: Query terms and techniques,’’ in Proc. 5th Annu. ACM Web Sci. Conf., 2013, pp. 195–204
[21] K. Reynolds, A. Kontostathis, and L. Edwards, ‘‘Using machine learning to detect cyberbullying,’’ in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA), Dec. 2011, pp. 241–244. VOLUME 7, 2019 70715 M. A. Al-Garadi et al.: Predicting Cyberbullying on Social Media.
[22] K. Dinakar, R. Reichart, and H. Lieberman, ‘‘Modeling the detection of textual cyberbullying,’’ in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 11–17.
[23] Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, ‘‘Detection of harassment on Web 2.0,’’ in Proc. Content Anal. Web, 2009, pp. 1–7.
[24] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, and K. Araki, ‘‘Machine learning and affect analysis against cyber-bullying,’’ in Proc. 36th AISB, 2010, pp. 7–16.
[25] Raisi and B. Huang, ‘‘Cyberbullying identification using participantvocabulary consistency,’’ 2016, arXiv:1606.08084. [Online]. 4
[26] N. Dang, M. Moreno-García and F. De la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study", Electronics, vol. 9, no. 3, p. 483, 2020. Available: 10.3390/electronics9030483 [Accessed 15 August 2020].
[27] V. Menger, F. Scheepers and M. Spruit, "Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text", Applied Sciences, vol. 8, no. 6, p. 981, 2018. Available: 10.3390/app8060981
[28] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, ‘‘Sarcasm as contrast between a positive sentiment and negative situation,’’ in Proc. Conf. Empirical Methods Natural Lang. Process., Oct. 2013, pp. 704–714
[29] M. Greenwood and D. Maynard, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis", in Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC, 2014, pp. 4238–4243.
[30] N. Mohamed Ali, M. El Hamid and A. Youssif, "sentiment analysis for movies reviews dataset using deep learning models", International Journal of Data Mining & Knowledge Management Process, vol. 09, no. 03, pp. 19-27, 2019. Available: 10.5121/ijdkp.2019.9302
[31] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text", Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544-2558, 2010. Available: 10.1002/asi.21416
[32] V. S. Chavan and S. S. Shylaja, ‘‘Machine learning approach for detection of cyber-aggressive comments by peers on social media network,’’ in Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI), Aug. 2015, pp. 2354–2358.
[33] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, ‘‘Improved cyberbullying detection using gender information,’’ in Proc. 25th DutchBelgian Inf. Retr. Workshop, 2012, pp. 1–3.
[34] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, ‘‘Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying,’’ in Proc. Int. Joint Conf. SOCO-CISIS-ICEUTE. Cham, Switzerland: Springer, 2014, pp. 419–428
[35] K. Dinakar, R. Reichart, and H. Lieberman, ‘‘Modeling the detection of textual cyberbullying,’’ in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 11–17.
[36] M. Al-Garadi et al., "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges", IEEE Access, vol. 7, pp. 70701-70718, 2019.
[37] Géron, Hands-on machine learning with Scikit-Learn and TensorFlow. O'Reilly Media, Inc., 2017.
[38] Singh, B. Kumar, L. Gaur and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification", 2019 International Conference on Automation, Computational and Technology Management (ICACTM), 2019. Available: 10.1109/icactm.2019.8776800
[39] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad and G. Choi, "GBSVM: Sentiment Classification from Unstructured Reviews Using Ensemble Classifier", Applied Sciences, vol. 10, no. 8, p. 2788, 2020.
[40] K. Reynolds, A. Kontostathis and L. Edwards, "Using Machine Learning to Detect Cyberbullying", 2011 10th International Conference on Machine Learning and Applications and Workshops, 2011. Available: 10.1109/icmla.2011.152
[41] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer and A. Mohammed, "Social Media Cyberbullying Detection using Machine Learning", International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, 2019.
[42] [7]M. Di Capua, E. Di Nardo and A. Petrosino, "Unsupervised cyber bullying detection in social networks", 2016 23rd International Conference on Pattern Recognition (ICPR), 2016. Available: 10.1109/icpr.2016.7899672
[43] Van Hee et al., "Automatic detection of cyberbullying in social media text", PLOS ONE, vol. 13, no. 10, p. e0203794, 2018. Available: 10.1371/journal.pone.0203794 [Accessed 26 October 2020].
[44] Zhang et al., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network", 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016.