CYBERBULLYING DETECTION IN SOCIAL NETWORK

[1]Edla Hareen

***Abstract*—The escalated usage of social networking sites and freedom of speech has given optimal ground to individuals across all demographics of cyberbullying and cyber aggression. This leaves drastic and noticeable impacts on the behavior of a victim, ranging from om disturbance in emotional well-being and isolation from society to more severe and deadly consequences. Automatic cyberbullying detection has remained a very challenging task since social media content is in natural language and is usually posted in unstructured free-text form, leaving behind the language norms, rules, and standards. Evidently, there exist a substantial number of research studies that primarily focus on discovering cyberbullying textual patterns over diverse social media platforms, as discussed previously in the literature review section. However, most of the detection schemes and automated approaches formulated are for resource-rich and mature languages spoken worldwide. English is commonly spoken around the world and is a language with limited resources. Hence, this research puts forth novel efforts to propose data pre-processing techniques on English scripting and develop deep learning-based hybrid models for automated cyberbullying detection in English. The outcomes of this study, if implemented, will assist cyber centers and investigation agencies in monitoring social content and making cyberspace a secure and safer place for all segments of society.**

***Index Terms*— CNN, DNN, Cyberbullying, Transfer learning**

I. OBJECTIVE

Social media has recently developed into one of the most influential communication tools and revolutionized millions of individuals globally. It has become an important tool in our daily lives and plays a significant role when it comes to expressing ourselves. Furthermore, virtual modes of communication have been developed and have helped corporations flourish and expand all over the world. People like to be recognized, and social media has turned out to be a paramount tool for them to express their states of being. Celebrities and other public figures use social media to share important information, such as their ideologies, and to update their fans and followers on their next move. Despite its positive effects on the world, social media has also led to and exacerbated bullying in society. It is a new form of bullying commonly referred to as "cyberbullying." Even though cyberbullying affects individuals in different ways, a clear correlation exists when it comes to self-esteem among victims and perpetrators. In the event that there are speculations and controversies involving particular people and organizations, they will use social media to express their stand or make comments to clear up the controversy or correct their status. However, the fact that social media is useful in our lives does not disqualify the drawbacks associated with social based on of cyberbullying. Notably, social media allows users to insult, bully, and threaten others without the fear of being punished, and most offenders believe that cyberbullying is not going to land them in any trouble. (Michael & Agur, 2018). This daring attitude has been fueled by the use of social media technology, which has instigated most of the cyberbullying cases. To combat behavior our, society should enact proper measures and policies that will be taken against the perpetrators. Without these policies, social media will continue to be a tool for body shaming and other cyberbullying acts, which can cause significant harm to victims.

II. MOTIVATION

Cyberbullying has ruined many lives. Cyberbullying on social media is linked to depression in teenagers, according to new research that analyzed multiple studies of the online phenomenon. Social media use is hugely common among teenagers, said Michele Hamm, a researcher in pediatrics at the University of Alberta, but the health effects of cyberbullying on social media sites are largely unknown. Both bullies and their victims are more likely to suffer from depression than youth who are not involved in bullying. This connection can be long-lasting; people who are bullied as children are more likely to suffer from depression as adults than children who are not involved in bullying. Teens who commit suicide often suffer from depression. Experts hesitate to say that bullying is a direct cause of suicide, but it may be a factor in a teen's depression. The Cyberbullying Research Center found that victims of cyberbullying were more likely to suffer from low self-esteem and suicidal thoughts. They suggest further research needs to be done to see if low self-esteem is a result of being cyberbullied or if it makes a person more likely to be a target of cyberbullying. A recent study by the US National Institutes of Health, reported by Reuters, found that victims of cyberbullying showed more signs of depression than other bullying victims. This may be because cyberbullying can be more relentless and more frightening or discouraging, especially if the bully is anonymous. Anxiety is also a huge factor when it comes to cyberbullying.

III. INTRODUCTION:

a) purpose: Cyberbullying can be tough to spot. Many young people who are being bullied don't want to tell teachers or parents, perhaps because they feel ashamed or they worry about losing their computer privileges at home. Parents often tell their children to turn off their mobile phones or stay off the computer. Many parents are unaware that the internet and mobile phones provide teenagers with a social lifeline to their peer group. Recently, there has been much media attention concerning this topic and its relationship to suicide. It is unknown whether other factors play a part, but cyberbullying is a contributing element in teen suicide. Many were affected by Sheniz Erkan's suicide, a victim of cyberbullying who was

sadly too afraid to speak up. Interestingly, a third of those who experience cyberbullying do not report it. If we are to succeed in preventing bullying, we need to break the climate of silence in which it thrives by empowering children and young people to speak out and seek help. Cyberbullying can have deleterious effects on a child's mental health. In particular, it can leave teenagers with low self-esteem, depression, anxiety, less interest in school, a deep sense of loneliness, self-harming, and, in some cases, suicide.

b) Scope: Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging, or digital messages. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. So to overcome these issues, detecting cyberbullying is very important these days, which will help to stop cyberbullying on social media networks.

IV. DETAILED DESCRIPTION ABOUT THE DATA STRUCTURE CONCEPT USED IN THE PROJECT:

DNN models learn word embeddings over the training data. These learned embeddings across multiple datasets show the difference in nature and style of bullying across cyberbullying topics and SMPs. Here we report results for BLSTM with an attention model. Results for other models are similar. We first verify that the important words for each topic of cyberbullying form clusters in the learned embeddings. To enable the visualization of grouping, we reduced dimensionality with t-SNE [8], a well-known technique for dimensionality reduction particularly well suited for the visualization of high-dimensional datasets. Each cluster shows that the words most relevant to a particular topic of bullying form a cluster. We also observed changes in the meanings of the words across topics of cyberbullying. The Twitter dataset, which is heavy on sexism and racism, considers the word "slave" as similar to targets of racism and sexism. The Wikipedia dataset on personal attacks, on the other hand, does not exhibit this bias. We used transfer learning to check if the knowledge gained by DNN models on one dataset could be used to improve cyberbullying detection performance on other datasets. We report results where BLSTM with attention is used as the DNN model. Results for other models are similar. We experimented with the following three flavors of transfer learning:

a) Complete Transfer Learning (TL1): In this flavor, a model trained on one dataset was directly used to detect cyberbullying in other datasets without any extra training. TL1 resulted in a significantly low recall, indicating that the three datasets have different natures of cyberbullying with low overlap. However, precision was relatively higher for TL1, indicating that DNN models are cautious in labeling a post as bullying. TL1 also aids in determining the degree of similarity in the nature of cyberbullying across three datasets. We can see that the nature of bullying is more similar in the Form spring and Wikipedia datasets than in the Twitter dataset. This can be inferred from the fact that with TL1, the cyberbullying detection performance for the Form spring dataset is higher when the base model is Wikipedia (precision = 0.51 and recall = 0.66) as compared to Twitter (precision = 0.38 and recall = 0.04). Similarly, for the Wikipedia dataset, Form spring acts as a better base model than Twitter while using the TL1 flavor of transfer learning. The nature of SMP might be a factor behind this similarity in the nature of cyberbullying. Both Form spring and Wikipedia are task-oriented social networks that allow anonymity and larger posts. Whereas communication on Twitter is short, free of anonymity, and not oriented towards a particular task.

b) Feature Level Transfer Learning (TL2): In this flavor, a model was trained on a single dataset, and only the learned word embeddings were transferred to another dataset to train a new model. As compared to TL1, the recall score improved dramatically with TL2. Improvements in precision were also significant. These improvements indicate that learned word embeddings are an essential part of knowledge transfer across datasets for cyberbullying detection.

c) Model Level Transfer Learning (TL3): In this flavor, a model was trained on one dataset, and learned word embeddings as well as network weights were transferred to another dataset for training a new model. TL3 does not result in any significant improvement over TL2. This lack of improvement indicates that the transfer of network weights is not essential for cyberbullying detection and that learned word embeddings are the key knowledge gained by the DNN models. DNN-based models coupled with transfer learning beat the best-known results for all three datasets. The previous best F1 scores for the Wikipedia and Twitter datasets were 0.68 and 0.93, respectively. We achieve F1 scores of 0.94 for both of these datasets using BLSTM with attention and feature-level transfer learning. For the Form spring dataset, authors have not reported an F1 score. Their method has an accuracy score of 78.5%. We achieve an F1 score of 0.95 with an accuracy score of 98% for the same dataset.

V. DETAILED DESCRIPTION OF METHODOLOGY:

a) Framework: Detection of cyberbullying on social media is a challenging task. The definition of what constitutes cyberbullying is quite subjective. For example, frequent use of swear words might be considered bullying by the general population. However, for teen-oriented social media platforms such as Form spring, this does not necessarily mean bullying (Table 2). Across multiple SMPs, cyberbullies attack victims on different topics such as race, religion, and gender. Depending on the topic of cyberbullying, vocabulary and the perceived meaning of words vary significantly across SMPs. For example, in our experiments, we found that for the word "fat," the most similar words, as per the Twitter dataset, are "female" and "woman" (Table 8). However, the other two datasets show no such bias against women. This platform-specific semantic similarity between words is a key aspect of cyberbullying detection across SMPs. The style of communication varies significantly across SMPs. For example, Twitter posts are short and lack anonymity. whereas poston Q&A-oriented SMPs

arelong and have the option of anonymity (Table 1). Fast-evolving words and hashtags in social media make it difficult to detect cyberbullying using simple filtering approaches based on swear word lists. The option of anonymity in certain social networks also makes it more difficult to identify cyberbullying because the bully's profile and history may not be available.

Past works on cyberbullying detection have at least one of the following three bottlenecks: First (bottleneck B1), they target only one particular social media platform. How these methods perform across other SMPs is unknown. Second (Bottleneck B2), they address only one topic of cyberbullying, such as racism and sexism. Depending on the topic, the vocabulary and nature of cyberbullying change. These models are not flexible enough to accommodate changes in the definition of cyberbullying. Third, they rely on handcrafted features such as a swear word list and POS tagging (Bottleneck B3). However, these handcrafted features are not robust against variations in writing style. In contrast to existing bottlenecks, this work targets three different types of social networks (Form spring: a Q&A forum, Twitter: microblogging, and Wikipedia: a collaborative knowledge repository) for three topics of cyberbullying (personal attack, racism, and sexism) without doing any explicit feature engineering by developing deep learning-based models along with transfer learninWe experimented with diverse traditional machine learning models (logistic regression, support vector machine, random forest, naive Bayes) and deep neural network models (CNN, LSTM, BLSTM, BLSTM with Attention) using a variety of representation methods for words (bag of character n-grammes, bag of word unigrams, GloVe embeddings, SSWE embeddings). A summary of our findings and research contributions is as follows:

- This is the first work that systematically analyses cyberbullying on various topics across multiple SMPs and applies transfer learning for a cyberbullying detection task.
- The presence of swear words is neither necessary nor sufficient for cyberbullying. Robust models for cyberbullying detection should not rely on such handcrafted features.
- Deep learning-based models outperform traditional machine learning models for cyberbullying detection tasks.
- Training datasets for cyberbullying detection contain only a few posts marked as bullying.
- This class imbalance problem can be tackled by oversampling the rare class.
- The vocabulary of words used for cyberbullying and their interpretation vary significantly across SMPs.

Table:

Dataset	# Posts	Classes	Length @95%	Max Length	Vocabulary size	Source
FormSpring	12k	2	62	1115	6058	[12]
Twitter	16k	3	26	38	5653	[16]
Wikipedia	100k	2	231	2846	55262	[18]

In the above-given table, we performed experiments using large, diverse, manually annotated, and publicly available datasets for cyberbullying detection in social media. We cover three different types of social networks: a teen-oriented Q&A forum (Form spring), a large microblogging platform (Twitter), and a collaborative knowledge repository (Wikipedia talk pages). Each dataset addresses a different topic of cyberbullying. The Twitter dataset contains examples of racism and sexism. The Wikipedia dataset contains examples of personal attack. However, the Form spring dataset is not specifically about any single topic. All three datasets have the problem of class imbalance, where posts labeled as cyberbullying are in the minority as compared to neutral posts. Variation in the number of posts across datasets also affects the vocabulary size, which represents the number of distinct words encountered in the dataset. We measure the size of a post in terms of the number of words in the post. For each dataset, there are only a few posts with large sizes. We truncate such large posts to the size of posts ranked at the 95 percentile in that dataset. For example, in the Wikipedia dataset, the largest post has 2846 words. However, the number of posts ranked at the 95 percentile in that dataset is only 231. Any post larger than size 231 in the Wikipedia dataset will be truncated by considering only the first 231 words. This truncation affects only a small minority of posts in each dataset. However, it is required for efficiently training various models in our experiments.

b) Existing Architecture: Cyberbullying has been recognized as a phenomenon at least since 2003. The use of social media exploded with the launch of multiple platforms such as Wikipedia (2001), My Space (2003), Orkut (2004), Facebook (2004), and Twitter (2005). By 2006, researchers had pointed out that cyberbullying was as serious a phenomenon as offline bullying. However, automatic detection of cyberbullying has only been addressed since 2009. As a research topic, cyberbullying detection is a text classification problem. Most existing works follow the same pattern: obtain a training dataset from a single SMP, engineer a variety of features with specific styles of cyberbullying as the target, employ a few traditional machine learning methods, and evaluate success in terms of F1 score and accuracy. These works heavily rely on handcrafted features, such as the use of swear words. These methods tend to have low precision for cyberbullying detection as handcrafted features are not robust against variations in bullying style across SMPs and bullying topics. Only recently has deep learning been applied to cyberbullying detection. Table 10 summarises important related work.

c) Proposed Architecture: Existing research has heavily relied on traditional machine learning models to detect cyberbullying. However, they do not study the performance of these models across multiple SMPs. We experimented with four models: logistic regression (LR), support vector machine (SVM), random forest (RF), and naive Bayes (NB), as these were used in previous works (Table 10). We used two data representation methods: character n-grammes and word unigrams. Past work in the domain of detecting abusive language has shown that simple n-gram features are more powerful than linguistic and syntactic features, hand-engineered lexicons, and word and paragraph embeddings. As compared to DNN models, the performance of all four traditional machine learning models was significantly lower. Please refer to the table below:

Dataset	label	Character n-grams				Word unigrams			
		LR	SVM	RF	NB	LR	SVM	RF	NB
Formspring	bully	0.448	0.422	0.298	0.359	0.489	0.463	0.264	0.025
Twitter	racism	0.723	0.676	0.752	0.686	0.738	0.772	0.739	0.617
	sexism	0.729	0.688	0.720	0.647	0.762	0.758	0.755	0.635
Wiki	Attack	0.694	0.677	0.674	0.655	0.711	0.686	0.730	0.659

All DNN models reported here were implemented using Keras. We preprocess the data, subjecting it to standard operations of removal of stop words, punctuation marks, and lowercasing, before annotating it to assigning respective labels to each comment. For each trained model, we report its performance after doing five-fold cross-validation. We use the following short forms.

- Datasets: F (Form spring), T (Twitter), W (Wikipedia).
- Datasets with oversampling of bullying posts: F+ (Form spring), T+ (Twitter),W+ (Wikipedia).
- Evaluation measures: P (precision), R (recall), F1 (F1 score).
- DNN models: M1 (CNN), M2 (LSTM), M3 (BLSTM), M4 (BLSTM with attention).

d) Module:

- Form spring: It was a question-and-answer-based website where users could openly invite others to ask and answer questions. There are 12K annotated question-and-answer pairs in the dataset. Each post is manually labeled by three workers. Among these pairs, 825 were labeled as containing cyberbullying content by at least two Amazon Mechanical Turk workers.
- Twitter: There are 16K annotated tweets in this dataset. The authors bootstrapped the corpus collection by performing an initial manual search of common slurs and terms used pertaining to religious, sexual, gender, and ethnic minorities. Of the 16K tweets, 3117 are labeled as sexist, 1937 as racist, and the remaining are marked as neither sexist nor racist.
- Wikipedia: For each page in Wikipedia, a corresponding talk page maintains the history of discussion among users who participated in its editing. This data set includes over 100,000 labeled discussion comments from English Wikipedia's talk pages. Each comment was labeled by 10 annotators via Crowd flower as to whether it contained a personal attack. There are 13590 comments classified as a personal attack.
- Use of Swear Words and Anonymity:

Dataset	P(B)	P(S)	P(A)	P(B|S)	P(S|B)	P(B|A)	P(A|B)	P(S|A)	P(B| (A&S))
FormSpring	0.06	0.16	0.53	0.22	0.59	0.08	0.71	0.20	0.25
Twitter	0.31	0.13	-	0.42	0.18	-	-	-	-
Wikipedia	0.12	0.17	0.27	0.49	0.69	0.25	0.56	0.27	0.65

B = bullying; S = swearing; A = anonymous. Some of the values in the Twitter dataset are undefined as Twitter does not allow anonymous postings. The use of swear words has been repeatedly linked to cyberbullying. However, preliminary dataset analysis reveals that depending on swear word usage, it cannot lead to high precision or recall for cyberbullying detection. Because P(B|S)

is not close to one, swear word list-based methods will have low precision. In fact, on the teen-oriented social network Form spring, 78% of the swearing posts are not bullying. Swear word-based filtering will be irritating to the users in such SMPs where swear words are used casually. Because P(S|B) is not close to one, swear word list-based methods will also have a low recall. For the Twitter dataset, 82% of bullying posts do not use any swear words. With swear word list-based methods, such passive-aggressive cyberbullying will go undetected. Anonymity is another clue that is used for detecting cyberbullying, as the bully might prefer to hide its identity. Anonymity increases the use of curse words (P(S|A) P(S) and cyberbullying (P(B|A) P(B) and P(B|(A&S)) P(B)). However, a significant fraction of anonymous posts are non-bullying (P (B|A) is not close to 1), and many of the bullying posts are not anonymous (P (A|B) is not close to 1). Further, anonymity might not be allowed by many SMPs, such as Twitter.

e) Deep Neural Network (DNN) Based Models: We experimented with four DNN-based models for cyberbullying detection: CNN, LSTM, BLSTM, and BLSTM with attention. These models are listed according to the increasing complexity of their neural architecture and the amount of information used by them. Please refer to the architecture in the existing general architecture architecture that we have used across four models different models differ only in the neural architecture layer, with the remaining layers being identical.CNN's are providing state-of-the-art results in extracting contextual features for classification tasks in images, videos, audio, and text. Recently, CNN's were used for sentiment classification. Long-Short Term Memory networks are a type of RNN that can learn long-term dependencies. Their ability to use their internal memory to process arbitrary sequences of inputs has been found to be effective for text classification. Bidirectional LSTMs further increase the amount of input information available to the network by encoding information in both forward and backward directions. By using two directions, input information from both the past and future of the current time frame can be used. Attention mechanisms allow for a more direct relationship between the state of the model at different points in time. Importantly, the attention mechanism lets the model learn what to attend to based on the input sentence and what it has produced so far. The embedding layer processes a fixed-size sequence of words. Each word is represented as a real-valued vector, also known as word embeddings. We have experimented with three methods for initializing word embeddings: random, GloVe, and SSWE. During the training, the model improves upon the initial word embeddings to learn task-specific word embeddings. We have observed that these task-specific word embeddings capture the SMP-specific and topic-specific styles of cyberbullying. Using GloVe vectors over random vector initialization has been reported to improve performance for some NLP tasks. Most word embedding methods, such as GloVe, consider only the syntactic context of the word while ignoring the sentiment conveyed by the text. The SSWE method overcomes this problem by incorporating text sentiment as one of the parameters for word embedding generation. We experimented with various dimension sizes for word embeddings. Experimental results reported here are for a dimension size of 50. With dimension sizes ranging from 30 to 200, there was no significant variation in results. To avoid overfitting, we used two dropout layers, one before the neural architecture layer and one after, with dropout rates of 0.25 and 0.5, respectively. The fully connected layer is a dense output layer with the number of neurons equal to the number of classes, followed by the soft max layer that provides soft max activation. Results for Traditional ML Models Using F1 Score

Dataset	label	Character n-grams				Word unigrams			
		LR	SVM	RF	NB	LR	SVM	RF	NB
Formspring	bully	0.448	0.422	0.298	0.359	0.489	0.463	0.264	0.025
Twitter	racism	0.723	0.676	0.752	0.686	0.738	0.772	0.739	0.617
	sexism	0.729	0.688	0.720	0.647	0.762	0.758	0.755	0.635
Wiki	Attack	0.694	0.677	0.674	0.655	0.711	0.686	0.730	0.659

f) Effect of Oversampling Bullying Instances: The training datasets had a major problem of class imbalance, with posts marked as bullying in the minority. As a result, all models were biased towards labeling the posts as non-bullying. To remove this bias, we oversampled the data from the bullying class three times. That is, we replicated bullying posts three times in the training data. This significantly improved the performance of all DNN models, with major leaps in all three evaluation measures. Table 4 shows the effect of oversampling for a variety of word embedding methods with BLSTM Attention as the detection model. Results for other models are similar. We can notice that oversampled datasets (F+, T+, and W+) have far better performance than their counterparts (F, T, and W, respectively). Oversampling particularly helps the smallest dataset, Form spring, where the number of training instances for the bullying class is quite small (825) as compared to the other two datasets (about 5K and 13K). We also experimented with varying the replication rate for bullying posts. However, we observed that for bullying posts, a replication rate of three is good enough.

Effect of Oversampling Bullying Posts using BLSTM with attention:

Dataset	Label	P			R			F1		
		Random	Glove	SSWE	Random	Glove	SSWE	Random	Glove	SSWE
F	bully	0.52	0.56	0.63	0.40	0.49	0.38	0.44	0.51	0.47
F+	bully	0.84	0.85	0.90	0.98	0.97	0.91	0.90	0.90	0.91
T	racism	0.67	0.74	0.76	0.73	0.76	0.77	0.70	0.75	0.76
T+	racism	0.94	0.90	0.90	0.98	0.95	0.96	0.96	0.93	0.93
T	sexism	0.65	0.86	0.83	0.64	0.52	0.47	0.65	0.65	0.59
T+	sexism	0.88	0.95	0.88	0.97	0.91	0.92	0.93	0.91	0.90
W	attack	0.77	0.81	0.82	0.74	0.67	0.68	0.76	0.74	0.74
W+	attack	0.81	0.86	0.87	0.91	0.89	0.86	0.88	0.88	0.87

g) Choice of Initial Word Embeddings and Model: Initial word embeddings determine data representation for DNN models. However, during training, DNN models modify these initial word embeddings to learn task-specific word embeddings. We have experimented with three methods to initialize word embeddings. Effect of Choosing Initial Word Embedding Method on F1 Score

Dataset	Label	Random		Glove		SSWE	
		M1	M4	M1	M4	M1	M4
F	bully	0.30	0.44	0.34	0.51	0.34	0.47
F+	bully	0.91	0.90	0.93	0.90	0.91	0.91
T	racism	0.68	0.70	0.73	0.75	0.70	0.76
T+	racism	0.90	0.96	0.95	0.93	0.93	0.93
T	sexism	0.59	0.65	0.61	0.65	0.63	0.59
T+	sexism	0.93	0.93	0.93	0.91	0.92	0.90
W	Attack	0.72	0.76	0.72	0.74	0.74	0.74
W+	Attack	0.83	0.88	0.89	0.88	0.88	0.87

This table shows the effect of varying initial word embeddings for multiple DNN models across datasets. We can notice that initial word embeddings do not have a significant effect on cyberbullying detection when oversampling of bullying posts is done (rows corresponding to F+, T+, and W+). In the absence of oversampling (rows corresponding to F, T, and W), there is a gap in the performance of the simplest (CNN) and most complex (BLSTM with attention) models. This gap, however, continues to close as dataset sizes grow larger.

h) Transfer Learning: We used transfer learning to check if the knowledge gained by DNN models on one dataset could be used to improve cyberbullying detection performance on other datasets. We report results where BLSTM with attention is used as the DNN model. Results for other models are similar. We experimented with the following three flavors of transfer learning:

- Complete Transfer Learning (TL1): In this flavor, a model trained on one dataset was directly used to detect cyberbullying in other datasets without any extra training. TL1 resulted in a significantly low recall, indicating that the three datasets have different natures of cyberbullying with low overlap.
- Feature Level Transfer Learning (TL2): In this flavor, a model was trained on a single dataset, and only the learned word embeddings were transferred to another dataset to train a new model. As compared to TL1, the recall score improved dramatically with TL2.

- Model Level Transfer Learning (TL3): In this flavor, a model was trained on one dataset, and learned word embeddings as well as network weights were transferred to another dataset for training a new model.

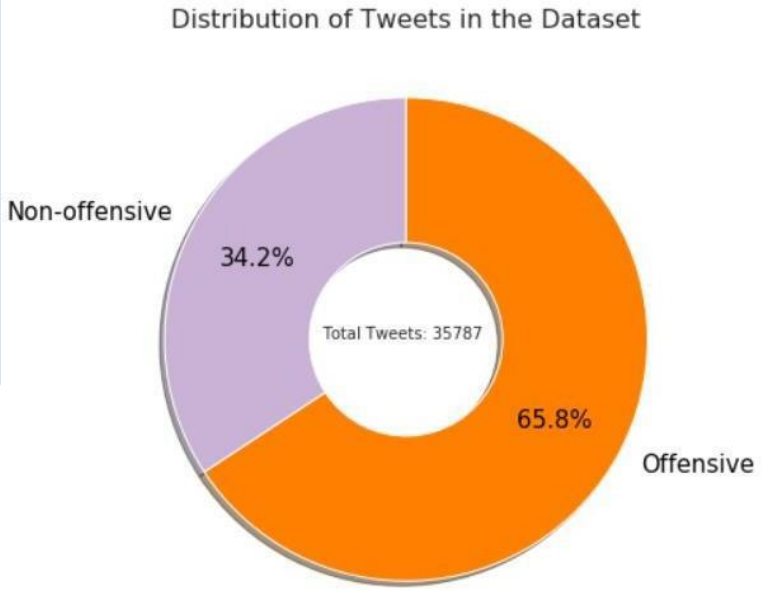
VI. Complete Demonstration of The Project:

This project aims to detect the offensive language in tweets using ML classification algorithms. A training and predicting pipeline is implemented to contrast the performance of various popular classification algorithms and determine the best-suited model. Data is taken from two sources:

I. Hate Speech Twitter Annotations:
- Publication: Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In NAACL SRW, pages 88–93, 2016
- Authors: Waseem, Zeerak and Hovy, Dirk
- Description: The dataset contains about 17,000 Tweet IDs labeled for racism and sexism. We downloaded this dataset and queried a Twitter API to scrape the actual tweets from Twitter. Retrieval of about tweets 5,900 failed either because the tweet was deleted or the account was deactivated.

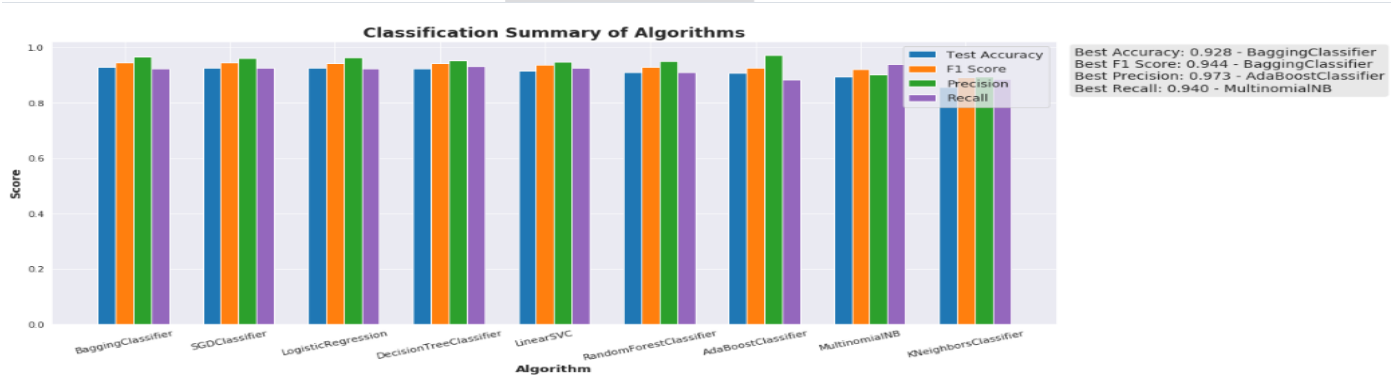
II. Hate Speech and Offensive Language Detection:
- Publication: Automated Hate Speech Detection and the Problem of Offensive Language.
- Authors: Davidson, Thomas, and Warmsley, Dana and Macy, Michael and Weber, Ingmar.
- Description: The dataset has about 25,000 Tweets annotated by crowd-sourcing. As per the number of users labeling the Tweets, each is put in one of three classes - hate speech, offensive language, and neither. We downloaded the dataset in Python from GitHub as a CSV file.

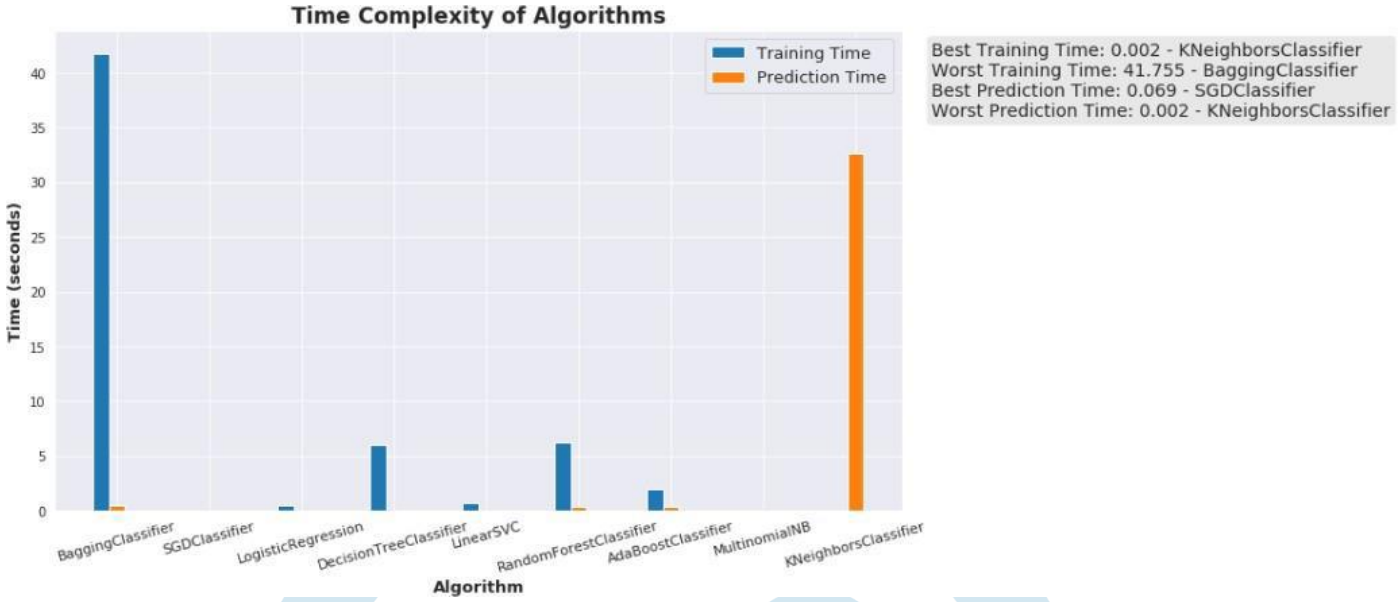
The data from both sources are clubbed. Here is the distribution of the tweets into the two classes:



The code is distributed into two Juypyter Notebooks which can be viewed in rendered format on the links:
- cyberbullying_wrangling.ipynb
- cyberbullying_v2.ipynb

VII. Summary:





After tuning hyper-parameters to optimize the algorithms, stochastic gradient descent was found to be the best-suited algorithm, taking both performance and time complexity into account. The following performance metrics were achieved:

- Accuracy: 92.81 %
- Precision: 96.97 %
- Recall: 91.94 %
- F1-Score: 94.39 %

VIII. CONCLUSION:

We have shown that DNN models can be used for cyberbullying detection on various topics across multiple SMPs using three datasets and four DNN models. These models, coupled with transfer learning, beat state-of-the-art results for all three datasets. These models can be further improved with extra data, such as information about the profile and social graphs of users. Most of the current datasets do not provide any information about the severity of bullying. If such fine-grained information is made available, then cyberbullying detection models can be further improved to take a variety of actions depending on the perceived seriousness of the posts.

REFERENCES

[1] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In WWW, pages 759–760, 2017.

[2] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In WWW, pages 29–30, 2015.

[3] S. Hinduja and J. W. Patchin. Bullying, cyberbullying, and suicide. Archives of suicide research, 14(3):206–221, 2010.

[4] R. Johnson and T. Zhang. Supervised and semi-supervised text categorization using lstm for region embeddings. In ICML, pages 526–534, 2016.

[5] D. Karthik, R. Roi, and L. Henry. Modeling the detection of textual cyberbullying. In Workshop on The Social Mobile Web, ICWSM, 2011.

[6] Y. Kim. Convolutional neural networks for sentence classification. In EMNLP, pages 1746–1751, 2014.

[7] L. v. d. Maaten and G. Hinton. Visualizing data using t-see. Journal of Machine Learning Research, 9(Nov):2579–2605, 2008.

[8] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In WWW, pages 145–153, 2016.

[9] J. W. Patchin and S. Hinduja. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth violence and juvenile justice, 4(2):148–169, 2006.

[10] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In EMNLP, pages 1532–1543, 2014.

[11] K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In ICMLA, pages 241–244, 2011.

[12] R. L. Servance. Cyberbullying, cyber-harassment, and the conflict between schools and the first amendment. Wisconsin Law Review, pages 12–13, 2003.

[13] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In ACL, pages 1555–1565, 2014.

[14] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection and prevention of cyberbullying. In Intl. Conf. Human and Social Analytics, pages 13–18, 2015