Comparative Study for Predicting the Severity of Cyberbullying Across Multiple Social Media Platforms

Akshita Aggarwal, Kavita Maurya, Anshima Chaudhary
Department of Computer Engineering
Netaji Subhas University of Technology
New Delhi, India

Abstract— **Initially, cyberbullying detection was mostly done by explicit reporting of inappropriate content by users on the internet, but victims usually are not in the mental space to report such instances, which ultimately go unnoticed by the social media websites. A need for automated methods of cyberbullying detection was realized but the current detection methods require massive improvements to correctly determine the extent of cyberbullying. Our paper aims to identify instances of cyberbullying across various social media platforms and classify them based on the level of severity of cyberbullying. We wish to provide a comparative study of various traditional machine learning and deep learning models that can be used for the above objective.**

We also wish to release 3 datasets that are annotated in 4 labels based on the severity of cyberbullying (none, low, medium, and high) to be used for future work by the interested person(s). Since young adults are very vulnerable to such incidents, the paper aims to work on social media platforms trending amongst the youth, like Reddit, Twitter, and Formspring.

Keywords—machine learning, natural language processing, cyberbullying, deep learning, social media

I. INTRODUCTION

Cyberbullying is a kind of harassment that occurs over digital mediums like mobiles, desktops, or tablets. Cyberbullying can be carried out via messages, chats, and apps, or online on social platforms or live/online gaming through which users can indulge together and share data. Cyberbullying includes blackmail, insult, curse/ exclusion, breaching someone else's privacy, abuse, sexual harassment, and/or defamation. Some instances harm the victim gravely and can be categorized as unlawful. The most common occurrences of cyberbullying take place at:

- Social Interaction Platforms, such as Instagram, Snapchat, Facebook and Twitter
- SMS called Text Message sent through our phones
- Instant Message (via devices, WhatsApp, Telegram, Slack, Discord and Facebook messenger services)

The detection of cyberbullying occurring on social media platforms is hard mainly because the meaning of cyberbullying can vary from person to person, especially while predicting the severity, what might be a case of extreme severity for one, might not be for others.

The objective of our paper is to identify instances of cyberbullying across various social media platforms (Twitter, Reddit, and Formspring) and classify them based on the level of severity. We aim to classify the instances into low, medium, and high levels of severity of cyberbullying. We wish to provide a comparative study of various traditional machine learning and deep learning models that can be used for the above objective.

Ex: Any statements containing abusive language, personal attack, or degrading a particular gender/religion will be classified as a severe case of bullying. We wish to use multiple datasets, taken from different social media websites, and feed them to various traditional and deep learning algorithms to determine the best methods for determining to cyberbully and provide a comparative report. Our main motive behind predicting the severity of cyberbullying is to ensure that high severity cases can be reported and immediate actions can be taken to rescue the victim.

Summarizing the tasks performed,

- We were successful in running 4 traditional ML models namely SVM, Naive Bayes, Random Forest and Logistic Regression (for char and word embedding) and 4 deep learning models namely CNN, LSTM, BLSTM-Attention and BLSTM (GloVe and SSWE Embedding) over 3 major social media platforms Twitter, Formspring and Reddit.
- We successfully labeled the data into suitable categories to the best of our understanding.
- We drew meaningful insights from the results obtained that could be useful in future research in this domain.

The paper is organized as follows – section II discusses the background and related work in this domain. Section III focuses on the datasets used for training and testing the models. Section IV enumerates on the methodology we used to perform the experiments. Section V describes the evaluation metrics used. Section VI contains details of the results obtained and section VII consists of the conclusions drawn from the results. We have discussed the limitations of the study in section VIII and future work in section IX.

978-1-7281-4876-2/20/$31.00 ©2020 IEEE

II. RELATED WORK

Detection of cyberbullying requires intelligent systems because it is difficult to understand the complexities involved with text classification. Many machine learning models have been developed so far but deep learning models have not been exploited to the fullest in this domain. Moreover, current ML models work fine on a single social media platform (SMP) but fail when the same model is used on a different SMP. There are three ways by which cyberbullying detection can be done[1]:-

1. Content-Based Cyberbullying Detection - There have been existing models that classify twitter data or tweets based on their positivity or negativity [2]. Post that, they are further classified as positive tweets with/without cyber bullying or negative tweets with/without cyberbullying. ML models like Naive Bayes gave good accuracy of 70% along with usage of TF-IDF and LDA for the understanding context of words [3]. These models didn't do well when it came to slang and swear words but later some models incorporated emoticons/emojis into the models, which paired with SVM gave an accuracy of 81% [4]. The textual features mostly used by prior models comprised of the density of swear words, uppercase alphabets, number of exclamation and question marks, number of emoticons used, and part-of-speech tags used in a tweet/post/comment. In [5] top 5 Arabic swear words were selected and the tweets containing these words were directly classified as cyberbullying by a Naive Bayes classifier with a 90% accuracy rate.

2. User-Based Cyberbullying Detection – Gathering information from user's profiles like age, several posts/comments/data they share could implicate potentiality to harm others. In [6] age was considered as an important feature along with a history of the user, and it was assumed that if a user was a victim of bullying in the past, the instance can be repeated. Features based on user information were also used in [7] but under the assumption that different genders and different age groups behave differently on social media. The user location was also used as a feature and was incorporated into the models. However, all these models worked on single SMP only because not all websites ask for age, location, or store history of a user, hence these models cannot be extended across multiple SMPs. For example, websites like Twitter ask for age, name, and verify the user, however, SMP like Tinder does not verify whether real or fake names are being used on their platform.

3. **Network-Based Cyberbullying Detection** – Since most SMPs are based on graphs, another way to detect cyberbullying is by analyzing the social structure of users. This is done by deriving relevant features from the relationship-graph of users. In [8], they derived features from the social network graph, by evaluating the in-degree (popularity) of the graph along with its out-degree (activity) since studies show that victims of bullying are more active on social media. The number of edges and nodes present in the relationship graph was used to infer the type and density of the community surrounding the user. Another approach visualized an account's profile based on the past bullying posts made by the user and checked whether the user is a victim/predator or a bystander (defendant or assistant).

However, the current automatic cyberbullying detection methods still need improvements to accurately identify the extent of cyberbullying in a particular post.

Issues with prior work

It is noticed that the prior work done in the field of detecting cyberbullying, has certain limitations(one or more) -

First (Limitation L1) – Determining Severity – although there are enough models to detect the presence and absence of cyberbullying, not much work has been put into determining the severity of bullying. By knowing the severity, it is possible to directly report the cases of extreme abuse/ bully, and appropriate actions can be taken by the concerned authorities. Lower levels of bullying can be dealt with by putting certain restrictions on the abuser, while higher levels must be subject to stringent restrictions and even permanent blockage of the online bullying user account, hence protecting the victim at an early stage.

Second (Limitation L2) – Performance across multiple SMPs – most cyberbullying detection platforms can work on a single type of social media platform because each one has its kind of content, hence there is an absence of a generalized model that can distinguish between cyberbullying and non-cyberbullying cases. Most models are run on only one social networking website, and cannot be used for other platforms. Knowing the behavior of users on multiple media platforms is essential to identify and the general trend and to profile the abuser.

Third (Limitation L3), – Dependency on user-determined features – current models rely on carefully user-determined parameters. However, these features are not robust against different kinds of social media platforms and for users of different kinds. For example, the use of swear words does not necessarily imply the presence of abuse, because swear words can be used in different contexts. For example, in SMPs which are teen-oriented or meant for young adults or comical, the use of abusive language might not necessarily imply bullying. To work on the same, word embedding techniques can be used to infer the contextual meaning of the phrases and words into consideration.

In this paper, we tried to work on these limitations since we focused on determining the severity of cyberbullying, used multiple datasets, and reduced the dependency on hand-crafted features such as swear words. We extended the work done in [9] and the features and models mentioned in it served as a basis for constructing our machine learning and deep learning models as well.

III. DATASETS

We labeled the dataset into 4 categories namely:

- None (No cyberbullying)
- Low cyberbullying (L)
- Medium cyberbullying(M)
- High cyberbullying(H)

We have used datasets from three different social platforms:

Formspring: A Q/A website where users (mainly teens or college students) interacted with each other by asking and answering questions[10]
Twitter: Popular microblogging where users post and interchange information via messages called "tweets".
Reddit: Reddit is a news accumulation website where users can rate web content and create sub-groups for discussions (called subreddits).
Major issues encountered with these datasets are:

- A large number of 'none' cases and fewer cases of cyberbullying. We used oversampling to solve this issue.
- Many posts were of variable length. We considered only a particular length of large posts. Although the downside of trimming down these posts is that it reduces its context and the cyberbullying might be present in the trimmed portion of such posts.
- Presence of recent short forms, slang, and emoticons (typographical or visual) for which the word embedding or representations weren't easily available.
- While labeling the data, cases were encountered where it was difficult to label a particular content as low or medium and there was also some uncertainty whether to label data as medium or high.

IV. PROPOSED METHODOLOGY

A. Collect data from multiple SMPs

The datasets of Formspring[11], Twitter[12] and Reddit[9] are taken to observe the performance of various ML/DL models. These datasets are diverse since each of the forums is used by different age groups and has a different type of content. These particular datasets were chosen because each dataset consisted of different forms of cyberbullying - dissing, online harassment, trolling, outing, or flaming.

- The data consisted of User ID and comments made by people on tweets, in the case of Twitter Dataset.
- The Formspring data comprised of user ID, question-answer type of format, between 2 users, chatting with each other.
- The Reddit Dataset also comprised of a user ID along with their comments.

The datasets of each SMP were collected from the sources mentioned in the research papers[9,12] and the large corpus of datasets at the Kaggle website[11] was referred to, from which the Formspring data was collected. In totality, the three datasets targeted the comment section of Twitter, Question-Answer format of Formspring, and the posts from Reddit platform.

B. Annotate the datasets

The dataset was categorized by giving each post/tweet/comment in the dataset one of the 4 labels – None(N), Low(L), Medium(M), and High(H) indicating the severity of bullying involved.

- Twitter data mostly comprised of racist and sexist comments by various people of different nationalities and older age brackets.

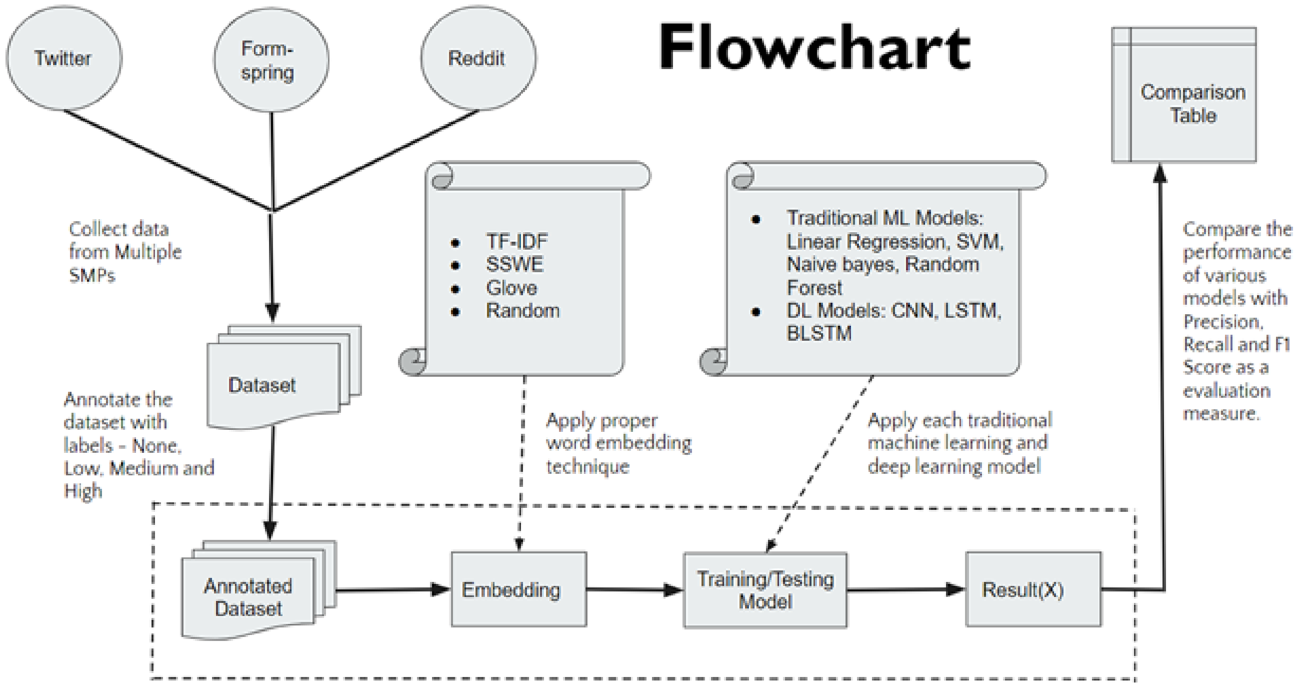

Twitter
Form-spring
Reddit
Flowchart
Comparison Table
Collect data from Multiple SMPs
TF-IDF
SSWE
Glove
Random
Traditional ML Models: Linear Regression, SVM, Naive bayes, Random Forest
DL Models: CNN, LSTM, BLSTM
Compare the performance of various models with Precision, Recall and F1 Score as a evaluation measure.
Dataset
Annotate the dataset with labels - None, Low, Medium and High
Apply proper word embedding technique
Apply each traditional machine learning and deep learning model
Annotated Dataset
Embedding
Training/Testing Model
Result(X)

Figure 1: Flowchart for the proposed methodology

- The Reddit data, on the other hand, comprised of younger users, hence it comprised of more lingos and slang, and unnecessary usage of swear words was also prevalent, which did not necessarily indicate cyberbullying, so care had to be taken while labeling that dataset.
- The Formspring data comprised of a question-answer format, which contained more instances of harassment, blackmail, dissing, and threats[10].

The annotation of data was done by the students after careful research and understanding of what falls into the category of cyberbullying and comprehending the severity of cyberbullying present in the posts. None was assigned to the data which had no trace of cyberbullying in it, while low was given to the data which consisted of mildly offensive or mildly inappropriate content, which can/cannot be taken as cyberbullying by the victim. The data consisting of threats, offensive content, and racist/sexist/dirty/abusive remarks targeting an individual or a community were adjudged as 'medium'. The instances where there were death threats and extremely intolerable comments about someone's physical looks or mental faculties or sexual harassment which made the comment completely unacceptable as per the society standards were all labeled as 'High.'

C. Identify appropriate word embedding

Being a text classification problem, the words of the dataset need to be first converted to vector representations. Word embedding is learned from unannotated plain text, useful in determining the context in which a given word is used[13]. They provide a dense vector representation of syntactic or semantic aspects of a word. **Various word embedding techniques are GloVe, SSWE, TF-IDF, Word2Vec, and Fast text. The word embedding chosen for the data was GloVe and SSWE.** These were chosen because being trained on a large corpus of words, they provided an edge over the other embedding. Also, as per our understanding, there were certain drawbacks of Fast text and Universal Sentence Encoder due to which GloVe and SSWE were chosen for implementing deep learning models.

- The reasons for choosing the GloVe word embedding over others are: - The goal of GloVe embedding is very straightforward, which is, to enforce the word vectors to discover the sub-linear word relationships in the complete vector search space. GloVe adds practical meaning to word vectors since it considers the relationships between two-word pairs, rather than between two words.
- The reasons for choosing SSWE are - SSWE encodes sentiment information by a representation of words that are continuous [14], which other word embeddings don't. When SSWE was applied to a Twitter sentiment classification dataset in [15] it was seen that the SSWE feature performs better by concatenating it with an existing feature set, hence making it relevant for Twitter-like datasets.

We used 3 different dimension sizes of word embedding – 50, 100, and 200. Based on the experimental results, the results for size 100 and 200 were not satisfactory [16]. After running multiple CNN models, we chose dimension size for word embedding as 50 and applied the same in the remaining models used in the study.

D. Apply traditional ML and DL models

- The traditional ML models, namely **- Logistic regression, support vector machines(SVM), Naive Bayes, random forest**, are used and their performances are compared. For the traditional ML Models, the Python libraries sklearn, and NumPy were used. Character and word-based embedding were used.
- The deep learning models used in the study are **CNN, LSTM, BLSTM, and BLSTM-A** which usually perform better than the traditional ML models. Tensorflow and Keras were used for running the deep learning models, using inbuilt libraries. SSWE and GloVe are used as word embedding.

E. Compare the performance of various algorithms

To compare the performance of algorithms, evaluation metrics like Precision(P), Recall(R), F1 score(F1) are used. **F1 Score** seems to be the best criterion for comparing various algorithms as it keeps both Precision and Recall into consideration. Hence most of the tabular results would be based on the use of F1 Score as the criteria for comparison**.** The detailed description, analysis, and conclusions derived from the experiments can be found in the consequent pages.

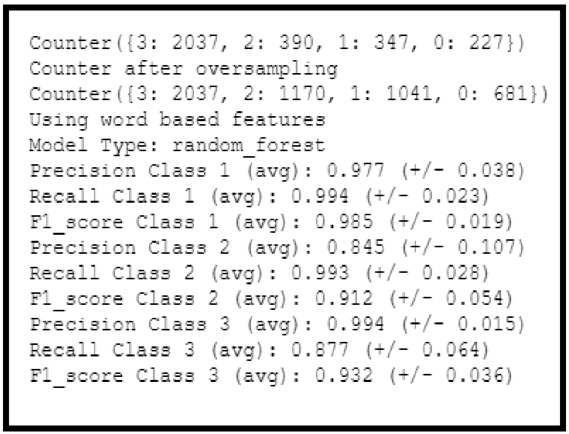
V. EVALUATION METRIC

The study aimed to:

- Highlight results of various algorithms in terms of performance matrices highlighting measures like F1 score, precision, recall, and accuracy.
- Tables showcase the exact results obtained from the algorithm. We have incorporated the following tables -
  - ✓ Comparison of F1 Score of various Traditional Models without oversampling for both character and word embedding.
  - ✓ Comparison of F1 Score of various Traditional Models with oversampling for both character and word embedding.
  - ✓ Comparison of F1 Score of various Deep Learning Models with oversampling for both SSWE and GloVe.
  - ✓ Comparison of accuracy for various Deep Learning Models for both SSWE and GloVe.
- Graphs showing various trends observed to draw meaningful conclusions

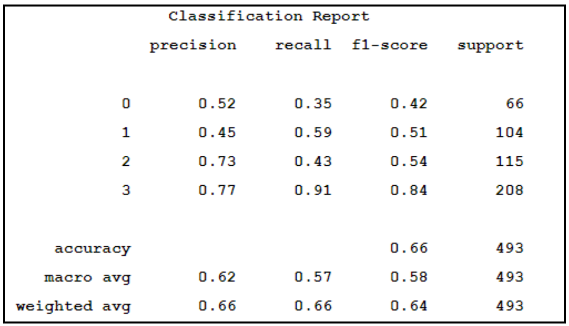
VI. RESULTS AND ANALYSIS

A. Classification Matrices of specific algorithmns



```
Counter({3: 2037, 2: 390, 1: 347, 0: 227})
Counter after oversampling
Counter({3: 2037, 2: 1170, 1: 1041, 0: 681})
Using word based features
Model Type: random_forest
Precision Class 1 (avg): 0.977 (+/- 0.038)
Recall Class 1 (avg): 0.994 (+/- 0.023)
F1_score Class 1 (avg): 0.985 (+/- 0.019)
Precision Class 2 (avg): 0.845 (+/- 0.107)
Recall Class 2 (avg): 0.993 (+/- 0.028)
F1_score Class 2 (avg): 0.912 (+/- 0.054)
Precision Class 3 (avg): 0.994 (+/- 0.015)
Recall Class 3 (avg): 0.877 (+/- 0.064)
F1_score Class 3 (avg): 0.932 (+/- 0.036)
```

Figure 2: Twitter Random forest word embedding



```
                    Classification Report
               precision    recall  f1-score   support

           0       0.52      0.35      0.42        66
           1       0.45      0.59      0.51       104
           2       0.73      0.43      0.54       115
           3       0.77      0.91      0.84       208

    accuracy                           0.66       493
   macro avg       0.62      0.57      0.58       493
weighted avg       0.66      0.66      0.64       493
```

Figure 3: Twitter CNN for SSWE embedding

B. Tables for comparison

1. Table 1:Comparison of F1 Score of various traditional models without oversampling for both char & word

Dataset	Label	Character Embedding				Word Embedding			
		LR	SVM	NB	RF	LR	SVM	NB	RF
Twitter	L	0.126	0.150	0.000	0.000	0.000	0.103	0.000	0.042
	M	0.278	0.302	0.000	0.093	0.257	0.275	0.000	0.249
	H	0.389	0.393	0.000	0.145	0.513	0.412	0.054	0.348
Formspring	L	0.064	0.172	0.000	0.030	0.062	0.232	0.000	0.096
	M	0.146	0.172	0.000	0.103	0.163	0.175	0.000	0.223
	H	0.531	0.593	0.000	0.398	0.537	0.510	0.000	0.551
Reddit	L	0.036	0.125	0.000	0.008	0.018	0.126	0.000	0.026
	M	0.016	0.089	0.000	0.008	0.020	0.097	0.000	0.006
	H	0.008	0.062	0.000	0.000	0.013	0.043	0.000	0.000

2. Table 2: Comparison of F1 Score of various traditional models with oversampling for both character & word

Dataset	Label	Character Embedding				Word Embedding			
		LR	SVM	NB	RF	LR	SVM	NB	RF
Twitter	L	0.537	0.780	0.011	0.999	0.910	0.976	0.650	0.994
	M	0.561	0.743	0.144	0.993	0.880	0.970	0.812	0.980
	H	0.636	0.783	0.319	0.990	0.919	0.967	0.855	0.915
Formspring	L	0.563	0.793	0.164	0.986	0.872	0.965	0.807	0.978
	M	0.664	0.877	0.034	0.982	0.950	0.973	0.606	0.979
	H	0.800	0.945	0.000	0.998	0.989	0.998	0.890	1.00
Reddit	L	0.397	0.481	0.033	0.989	0.807	0.882	0.978	0.914
	M	0.314	0.420	0.010	0.998	0.835	0.907	0.530	0.97
	H	0.299	0.378	0.009	0.996	0.856	0.909	0.206	0.972

3. Table 3: Comparison of F1 Score of various Deep Learning Models with oversampling for both SSWE and GloVe.

Dataset	Label	SSWE				GLOVE			
		LSTM	CNN	BLSTM	BLSTM-A	LSTM	CNN	BLSTM	BLSTM-A
Twitter	L	0.75	0.42	0.90	0.73	0.66	0.64	0.87	0.78
	M	0.07	0.51	0.87	0.72	0.76	0.55	0.85	0.82
	H	0.67	0.54	0.90	0.84	0.89	0.74	0.91	0.90
Form spring	L	0.00	0.39	0.60	0.35	0.00	0.67	0.55	0.62
	M	0.21	0.00	0.48	0.21	0.00	0.00	0.00	0.12
	H	0.37	0.53	0.71	0.72	0.38	0.67	0.73	0.87

4. Table 4: Comparison of accuracy for various Deep Learning Models for both SSWE and GloVe.

Dataset	SSWE				GLOVE			
	LSTM	CNN	BLSTM	BLSTM-A	LSTM	CNN	BLSTM	BLSTM-A
Twitter	64	68	67	67	67	69	69	69
Twitter +	72	66	91	84	83	76	90	87
Formspring	80	83	82	81	80	82	82	82
Form spring +	55	65	75	68	57	79	75	77

C. Graphs for observing trends

1. Comparison of F1 Score, Precision and Recall for all traditional models like Random Forest, Naïve Bayes, Logistic Regression and SVM (with oversampling for word embedding)

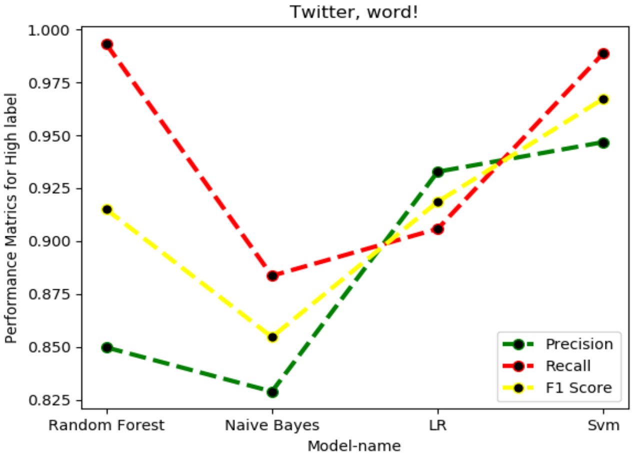


Figure 4: Graph for traditional models comparison

2. Comparison of F1 Score of all Deep Learning models like LSTM, CNN, BLSTM, and BLSTM-A
(With oversampling for GloVe embedding)

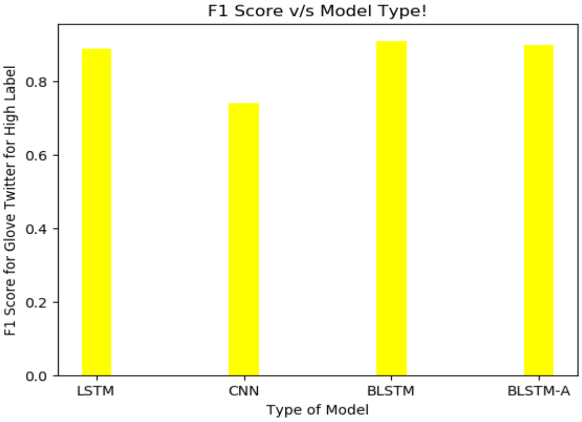


Figure 5: Graph for deep learning model comparison

VII. CONCLUSION

A. Traditional Models

- Without oversampling, the results of traditional models are extremely poor.
- With oversampling, the results of models have improved but we sense some degree of overfitting.
- In order of performance, SVM > Random Forest >= Logistic regression > Naive Bayes.
- Naive Bayes' performance is below acceptable levels and we recommend not using this algorithm at all.
- Word embedding seems to perform better when compared to char embedding.

B. Deep Learning Models

- Without oversampling, the results of deep learning models are extremely poor. This is mainly attributed to lesser instances of cyberbullying cases in any available datasets.
- With oversampling, the results of models have improved but owing to fewer data, results aren't as expected and observed from other sources.
- In order of performance for GloVe embedding, BLSTM > BLSTM-A > LSTM >CNN.
- In order of performance for SSWE embedding, BLSTM> BLSTM-A > LSTM >CNN.
- GloVe embedding seems to perform much better when compared to SSWE embedding.
- Accuracy doesn't seem to be good enough criteria for comparison of classification results. We have preferred the F1 Score for the portrayal of model comparisons. Accuracy decreases for some models when oversampling is done. Though other classification matrices show a significant increase. Hence, we disregard the use of accuracy as the real criteria for comparison.

C. Datasets specific observations

- Reddit doesn't give satisfactory results. Our assumption about the same is that our model isn't able to correctly identify instances of bullying, because in Reddit non-bullying comments also have to swear words. Deep learning models perform well below satisfactory and hence their results for Reddit have been excluded.
- Twitter gives the best results because it is the most appropriately labeled dataset as a lot of past work has been done concerning this dataset.

D. Relative Comparison between Traditional and Deep learning models

- Our first assumption would be that given their complexities, deep learning models would perform better than traditional.
- To our extremely astonishing conclusions, traditional models performed much better than deep learning on almost all classification matrices like precision, F1 Score, and Recall.
- The above result is difficult to perceive and can be attributed to several reasons. We believe that overfitting in traditional and lack of huge data for deep learning could be a reason for this stark difference in performance.
- But we wish to make an extremely important point here. The reason most cyberbullying analysis still focuses on using traditional models especially SVM is that on average they perform as good if not better than deep learning with fewer data and fewer resources.

The results are largely based on our assumptions and can be prone to many errors.

VIII. LIMITATIONS OF THE STUDY

- As the data has been labeled manually by our team, there are chances of the introduction of human bias in the results. What might be a case of extreme bullying for one, might not be for the other.
- Lack of huge corpus of data because of our limited ability to label data in a limited period. We believe the reason for lower than expected results for deep learning models is because of this issue only.

IX. FUTURE WORK

- A major way the model can do significant advances is by profiling the users. The availability of information of users (their age, gender, region, profession) can help us draw many valuable insights about both the abuser and the victim. This will enable us to ban any such accounts beforehand that pose a potential danger of causing cyber-abuse.

- Also, we believe another area of work could be the transfer learning between models where results of one dataset can be fed to another, and impacts can be seen.
- Increasing the size of data and taking help from experts while labeling can also be of great future significance.

POTENTIAL BENEFICIARIES

We believe that social media sites that aim at tracking cyber-bullying and wanting the high severity cases to be reported automatically are the main beneficiaries. Cyberbullying is a major challenge to modern-day social media and tracking that down can have huge positive impacts. We have annotated 3 datasets – Twitter, Reddit, and Formspring into 4 categories – (None, Low, Medium, and High) based on the severity of bullying, which can be utilized for future work. It can be found on the GitHub repository - https://github.com/Kavita309/Severity-of-cyberbullying-across-SMPs.

ACKNOWLEDGMENT

We wish to express our deepest gratitude towards Dr. Pinaki Chakraborty, Assistant Professor, Department of Computer Engineering, Netaji Subhas University of Technology, New Delhi, for his advice during this study.

REFERENCES

[1] International Journal of Advanced Computer Science and Applications,(IJACSA) Vol. 9, No. 9, 2018 - Deep Learning Algorithm for Cyberbullying Detection by Monirah Abdullah Al-Ajlan and Mourad Ykhlef, University of Riyadh, Saudi Arabia

[2] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of International Conference on Computational Linguistics, pages 36–44.

[3] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," Commun. Inf. Sci. Manag. Eng., vol. 3, no. 5, p. 238, 2013.

[4] Paper, "Methods for detection of cyberbullying : A survey," no. October 2016.

[5] 4 E. A. Abozinadah, A. V Mbaziira, and J. H. J. Jr, "Detection of Abusive Accounts with Arabic Tweets," vol. 1, no. 2, 2015.

[6] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," pp. 2–5.

[7] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in Australasian Database Conference, 2014, pp. 160–171.

[8] P. K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis," pp. 3–6, 2014.

[9] Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms Sweta Agrawal, Amit Awekar Indian Institute of Technology, Guwahati, awekar@iitg.ernet.in

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

[11] Datasets from Kaggle: https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullying-detection

[12] 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

[13] Sentiment Embeddings with Applications to Sentiment Analysis, VL - 28, DO - 10.1109/TKDE.2015.2489653, IEEE Transactions on Knowledge and Data Engineering

[14] Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification∗ Duyu Tang† , Furu Wei‡ , Nan Yang\ , Ming Zhou‡ , Ting Liu† , Bing Qin† †Research Center for Social Computing and Information Retrieval Harbin Institute of Technology, China ‡Microsoft Research, Beijing, China \University of Science and Technology of China, Hefei, China

[15] Coooolll: A Deep Learning System for Twitter Sentiment Classification∗ Duyu Tang† , Furu Wei‡ , Bing Qin† , Ting Liu† , Ming Zhou‡ †Research Center for Social Computing and Information Retrieval Harbin Institute of Technology, China ‡Microsoft Research, Beijing, China

[16] Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study Maral Dadvar and Kai Eckert Web-based Information Systems and Services, Stuttgart Media University Nebenstrasse 8, 70569 Stuttgart, Germany