Detection of Cyberbullying on Social Media Using Machine Learning

Asha Akula[1], G.V. Ramana[2]

[1]*PG Student in Dept of CSE in Sree Vahini Institute of Science and Technology By-Pass Road, India*
[2]*Associate Professor in Dept of CSE in Sree Vahini Institute of Science and Technology By-Pass Road, India*

Abstract: In this work, there is an argue for a focus on the latter problem for practical reasons. This project show that it is a much more challenging task, as the analysis of the language in the typical datasets shows that hate speech lacks unique, discriminative features and therefore is found in the 'long tail' in a dataset that is difficult to discover. Later in this project there is an propose of Deep Neural Network structures serving as feature extractors that are particularly effective for capturing the semantics of hate speech. These methods are evaluated on the largest collection of hate speech datasets based on Twitter, and are shown to be able to outperform state of the art by up to 6 percentage points in macro-average F1, or 9 percentage points in the more challenging case of identifying hateful content.

1. Introduction

The exponential growth of social media such as Twitter and community forums has revolutionised communication and content publishing but is also increasingly exploited for the propagation of hate speech and the organisation of hate-based activities [1, 3]. The anonymity and mobility afforded by such media has made the breeding and spread of hate speech eventually leading to hate crime effortless in a virtual land scape beyond the realms of traditional law enforcement. The term 'hate speech' was formally defined as 'any communication that disparages a person or a group on the basis of some characteristics (to be referred to as types of hate or hate classes) such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics. In the UK, there has been significant increase of hate speech towards the migrant and Muslim communities following recent events including leaving the EU, the Manchester and the London attacks. In the EU, surveys and reports focusing on young people in the EEA (European Economic Area) region show rising hate speech. And related crimes based on religious beliefs, ethnicity, sexual orientation or gender, as 80% of respondents have encountered hate speech online and 40% felt attacked or threatened. Statistics also show that in the US, hate speech and crime is on the rise since the Trump election. The urgency of this matter has been increasingly recognised, as a range of international initiatives have been launched towards the qualification of the problems and the development of countermeasures. Cyberbullying or cyberharassment is a form of bullying or harassment using electronic means. Cyberbullying and cyberharassment are also known as online bullying. It has become increasingly common, especially among teenagers, as the digital sphere has expanded and technology has advanced. Cyberbullying is when someone, typically a teenager, bullies or harasses others on the internet and in other digital spaces, particularly on social media sites. Harmful bullying behaviour can include posting rumors, threats, sexual remarks, a victims' personal information, or pejorative labels (i.e. hate speech). Bullying or harassment can be identified by repeated behaviour and an intent to harm. Victims of cyberbullying may experience lower self-esteem, increased suicidal ideation, and a variety of negative emotional 2 responses including being scared, frustrated, angry, or depressed. Cyberbullying can take place on social media sites such as Facebook, Myspace, and Twitter. "By 2008, 93% of young people between the ages of 12 and 17 were online. In fact, youth spend more time with media than any single other activity besides sleeping." The last decade has witnessed a surge of cyberbullying, which is categorized as bullying that occurs through the use of electronic communication technologies, such as e-mail, instant

messaging, social media, online gaming, or through digital messages or images sent to a cellular phone. There are many risks attached to social media sites, and cyberbullying is one of the larger risks. One million children were harassed, threatened or subjected to other forms of cyberbullying on Facebook during the past year, while 90 percent of social media-using teens who have witnessed online cruelty say they have ignored mean behaviour on social media, and 35 percent have done so frequently. Ninety-five percent of social-media-using teens who have witnessed cruel behaviour on social networking sites say they have seen others ignoring the mean behaviour, and 55 percent have witnessed this frequently. Terms like "Facebook depression" have been coined specifically in regard to the result of extended social media use, with cyberbullying playing a large part in this.

2. LITERATURE SURVEY

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Moreover, the response on twitter is prompterand also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis). Sentiment analysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analysing overall public sentiment towards that firm with respect to time and using economics tools for finding the correlation between public sentiment and the firm's stock market value. Firms can also estimate how well their product is responding in the market, which areas of the market is it having a favourable response and in which a negative response (since twitter allows us to download stream of geo-tagged tweets for particular locations. If firms can get this information, they can analyse the reasons behind geographically differentiated response, and so they can market their product in a more optimized manner by looking for appropriate solutions like creating suitable market segments. Predicting the results of popular political elections and polls is also an emerging application to sentiment analysis. One such study was conducted by Tumasjan et al. in Germany for predicting the outcome of federal elections in which concluded that twitter is a good reflection of offline sentiment.
[1] In afsaneh Asaei et al, Perceptual Information Loss because of Impaired Speech Production, Phonological classes characterize without articulatory and articulatorybound telephone properties. Profound neural system is utilized to gauge the likelihood of phonological classes from the discourse signal. In principle, a one of a kind mix of telephone characteristics structure a phoneme personality. Probabilistic induction of phonological classes in this manner empowers estimation of their compositional phoneme probabilities. An epic data theoretic system is concocted to measure the data passed on by each telephone trait, and survey the discourse creation quality for view of phonemes. As an utilization case, we theorize that interruption in discourse creation prompts data misfortune in telephone properties, and in this manner disarray in phoneme recognizable proof. We evaluate the measure of data misfortune because of dysarthria enunciation recorded in the TORGO database. A tale data measure is figured to assess the deviation from a perfect telephone credit creation driving us to recognize solid creation from obsessive discourse. [2] duanpei, m.tanaka and R.chen et al, a robust speech detection algorithm for speech activated hands-on application, depicts a novel commotion vigorous discourse discovery calculation that can work dependably in serious vehicle boisterous conditions. Superior has been acquired with the accompanying methods: (1) clamor concealment dependent on head part examination for pre-handling, (2) vigorous endpoint identification utilizing dynamic parameters [I] and (3) discourse check utilizing periodicity of voiced signs with symphonious improvement. Clamor concealment improves the SNR as contrasted and nonlinear range subtraction by around 20 db. This causes the endpoint location to work dependably in SNRs down to - 10 dB. In vehicle situations, street knock clamors are tricky for discourse identifiers causing mis-discovery blunders. Discourse confirmation assists with evacuating these blunders. This innovation is being utilized in Sony vehicle route items.

3. EXISTING SYSTEM

Existing methods primarily cast the problem as a supervised document classification task [33]. These can be divided into two categories: one relies on manual feature engineering that are then consumed by algorithms such as SVM, Naive Bayes, and Logistic Regression [3, 9, 11, 15, 19, 23, 35–39] (classic

methods); the other represents the more recent deep learning paradigm that employs neural networks to automatically learn multi-layers of abstract features from raw data [13, 26, 30, 34] (deep learning methods).The existing system lacks in calculating values with algorithms. These algorithms provide less accuracy.in existing system less data is tested if more data is tested then there will be an operational issue and provides less accuracy.

DISADVANTAGES OF EXISTING SYSTEM: • Existing studies on hate speech detection have primarily reported their results using micro-average Precision, Recall and F1 [1, 13, 30, 36, 37, 40]. • The problem with this is that in an unbalanced dataset where instances of one class (to be called the 'dominant class's) significantly out-number others (to be called 'minority classes'), micro-averaging can mask the real performance on minority classes.

4. PROPOSED SYSTEM

All datasets are significantly biased towards non-hate, as hate Tweets account between only 5.8% (DT) and 31.6% (WZ). When we inspect specific types of hate, some can be even scarcer, such as 'racism' and as mentioned before, the extreme case of 'both'. This has two implications. First, an evaluation measure such as the micro F1 that looks at a system's performance on the entire dataset regardless of class difference can be biased to the system's ability of detecting 'nonhate'. In other words, a hypothetical system that achieves almost perfect F1 in identifying 'racism' Tweets can still be overshadowed by its poor F1 in identifying 'non-hate', and vice versa. Second, compared to non-hate, the training data for hate Tweets are very scarce. This may not be an issue that is easy to address as it seems, since the datasets are collected from Twitter and reflect the real nature of data imbalance in this domain. Thus to annotate more training data for hateful content we will almost certainly have to spend significantly more effort annotating non-hate.

OBJECTIVE: This aims to classify textual content into non-hate or hate speech, in which case the method may also identify the targeting characteristics (i.e., types of hate, such as race, and religion) in the hate speech.

PROBLEM STATEMENT: Twitter is a popular social networking website where members create and interact with messages known as "tweets". This serves as a mean for individuals to express their thoughts or feelings about different subjects. Various different parties such as consumers and marketers have done sentiment analysis on such tweets to gather insights into products or to conduct market analysis. Furthermore, with the recent advancements in machine learning algorithms, we are able improve the accuracy of our sentiment analysis predictions. In this report, we will attempt to conduct sentiment analysis on "tweets" using various different machine learning algorithms. We attempt to 3 classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked as the final label. We use the dataset from Kaggle which was crawled and labelled positive/negative. The data pro-voided comes with emoticons, usernames and hashtags which are required to be processed and converted into a standard form. We also need to extract useful features from the text such uni-grams and bigrams which is a form of representation of the "tweet". We use various machine learning algorithms to conduct sentiment analysis using the extracted features. However, just relying on individual models did not give a high accuracy so we pick the top few models to generate a model ensemble. Assembling is a form of meta learning algorithm technique where we combine different classifiers in order to improve the prediction accuracy. Finally, we report our experimental results and findings at the end.

ADVANTAGE OF PROPOSED SYSTEM: • Also, as we shall show in the following, this problem may not be easily mitigated by conventional methods of over- or under-sampling. • Because the real challenge is the lack of unique, discriminative linguistic characteristics in hate Tweets compared to non-hate. • As a proxy to quantify and compare the linguistic characteristics of hate and non-hate Tweets, we propose to study the 'uniqueness' of the vocabulary for each class.

PSEUDO CODE FOR PROPOSED METHOD

```
BEGIN

1. IMPORT necessary libraries
  - Data handling: pandas, numpy
  - Text processing: re, nltk, sklearn
  - ML models: sklearn classifiers (e.g., Logistic Regression, SVM)

2. LOAD social media dataset
  - Data contains: post_id, user_id, text_content, label (bullying / not_bullying)

3. PREPROCESS the text data
  FOR each post in text_content:
    - Convert to lowercase
    - Remove URLs, hashtags, mentions, numbers
    - Remove punctuation
    - Tokenize text
    - Remove stop words
    - Lemmatize or stem words

4. EXTRACT features from text
  - Use TF-IDF or Count Vectorizer
  - Optionally add sentiment scores or keyword features

5. SPLIT dataset into training and testing sets (e.g., 80% train, 20% test)

6. SELECT and TRAIN a machine learning model
  - Examples: Naive Bayes, Logistic Regression, SVM, Random Forest
  - Fit model on training data

7. EVALUATE model on test data
  - Predict labels for test data
  - Compute accuracy, precision, recall, F1-score
  - Display confusion matrix

8. IF performance is acceptable THEN
     SAVE trained model for future use
  ELSE
     TRY other models or tune hyperparameters

9. FOR new social media post:
     - PREPROCESS the text
     - EXTRACT features
     - PREDICT using trained model
     - RETURN predicted label (bullying / not_bullying)

END
```

5. RESULTS

	precision	recall	f1-score	support
0	0.56	0.12	0.19	164
1	0.90	0.97	0.93	1905
2	0.84	0.81	0.83	410
accuracy			0.88	2479
macro avg	0.77	0.63	0.65	2479
weighted avg	0.87	0.88	0.86	2479

Accuracy Score: 0.8846308995562727

Fig: The Data Is Tested Under Bi Gram and The Accuracy Has Been Calculated

	precision	recall	f1-score	support
0	0.59	0.13	0.22	164
1	0.91	0.97	0.94	1905
2	0.84	0.85	0.84	410
accuracy			0.89	2479
macro avg	0.78	0.65	0.67	2479
weighted avg	0.88	0.89	0.87	2479

Accuracy Score: 0.891085114965712

Fig: The Data Is Tested Under Td Idf Mode and the Accuracy Score Is Calculated and Also Macro Avg, Weighted Avg.

	Neg	Pos	Neu	Compound	url_tag	mention_tag	hash_tag
0	0.000	0.120	0.880	0.4563	0.0	1.0	0.0
1	0.237	0.000	0.763	-0.6876	0.0	1.0	0.0
2	0.538	0.000	0.462	-0.9550	0.0	2.0	0.0
3	0.000	0.344	0.656	0.5673	0.0	2.0	0.0
4	0.109	0.229	0.662	0.6331	0.0	1.0	1.0
...	...	...	...	...	...	...	...
24778	0.000	0.000	1.000	0.0000	0.0	3.0	3.0
24779	0.454	0.000	0.546	-0.8074	0.0	0.0	0.0
24780	0.000	0.219	0.781	0.4738	0.0	0.0	0.0
24781	0.573	0.000	0.427	-0.7717	0.0	0.0	0.0
24782	0.000	0.218	0.782	0.5994	1.0	0.0	0.0

24783 rows × 7 columns

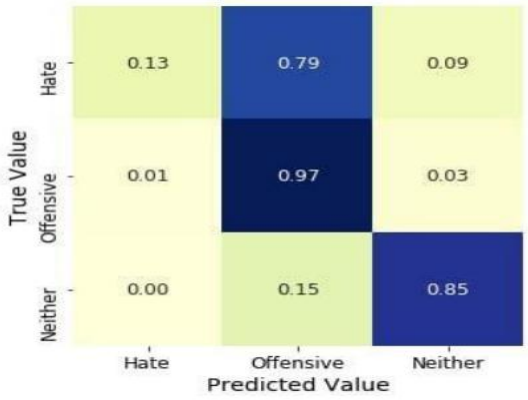

Fig: This is the final output in which the true values and predicted values are denoted. The values denote the percentage of classified words of hate, offensive, neither.

6. CONCLUSION

As hate speech continues to be a societal problem, the need for automatic hate speech detection systems becomes more apparent. In this report, we proposed a solution to the detection of hate speech and offensive language on Twitter through machine learning using Bag of Words and TF IDF values. We performed comparative analysis of Logistic Regression, Naive Bayes, Decision Tree, Random Forest and Gradient Boosting on various sets of feature values and model parameters. The results showed that Logistic Regression performs comparatively better with the TF IDF approach. We presented the current problems for this task and our system that achieves reasonable accuracy (89%) as well as recall (84%). Given all the challenges that remain, there is a need for more research on this problem, including both technical and practical matters.

FUTURE WORK: Generate new future's like removing those cyberbullying tweets that are posted in social media. increase more accuracy in prediction and warring the user about their tweets. We believe there are ways that design can help stop online aggression. Adding live detection or printing*** for bullying words can help in reduce of bullying or online distress. Better feedback from sites can encourage users to report aggression or harassment. Finally, existing designs can help support low-risk interventions. As we've seen, designing to help bystanders takes careful planning. It also requires sensitivity for the ways people use social media. Still, there's no shortage of ways to empower bystanders to stand up against online bullying.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27, 2011.

[3] G. Chowdhury. Introduction to modern information retrieval. Facet publishing, 2010.

[4] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In The Social Mobile Web, 2011.

[5] M. Fekkes, F. I. Pijpers, A. M. Fredriks, T. Vogels, and S. P. Verloove-Vanhorick. Do bullied children get ill, or do ill children get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms. Pediatrics, 117(5):1568–1574, 2006.

[6] G. Gini and T. Pozzoli. Association between bullying and psychosomatic problems: A meta-analysis. Pediatrics, 123(3):1059–1065, 2009.

[7] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle. Named entity recognition for twitter microposts using distributed word representations. In Proceedings of the Workshop on Noisy User-generated Text, pages 146–153, Beijing, China, July 2015. Association for Computational Linguistics.

[8] J. Juvonen and E. F. Gross. Extending the school grounds?ˆaA˘Tbullying experiences in cyberspace. ˇ Journal of School health, 78(9):496–505, 2008.

[9] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. 2014.

[10] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. Discourse processes, 25(2-3):259–284, 1998