CYBERBULLING DETECTION MODEL

MAYANK KAUSHIK

BTECH STUDENT, IT DEPARTMENT, MAHARAJA AGRASEN INSTITUTE OF TECHNOLOGY, DELHI, INDIA

---***---

Abstract – *Harassment of social media users by cyberbullies is a significant yet harmful phenomenon on the social media. Existing methods to detect cyberbullying have at least one of these following three bottlenecks. First they only target one social media platform, second is that they address only one topic of cyberbullying. Third is that they rely on carefully handcrafted features of the data. We in this project show that machine learning models can overcome all these three bottlenecks. Knowledge that is learned by these models on one type of dataset can be transferred and used to other datasets. Our experiments provide some very useful insights about cyberbullying detection. According to the best of our knowledge, this is work that systematically analyzes cyberbullying detection on various topics across various social media platforms using Machine Learning and Transfer learning*

Key Words: Cyberbullying, Social Media, Machine Learning

1. INTRODUCTION

As the number of users on social media are increasing. It leads to a new way of bullying. The later term is defined as an intentional or an aggressive acts that are carried out by person or groups of individuals using repeatedly communication messages over time against a victim or user who cannot easily defend him or herself [1]. Bullying has always been a part of our society since a long time. With change coming with internet, it was only a matter of time that bullies found their view on this new and opportunistic medium. Bullies has become able to do their nasty deeds with anonymity and great distance between them and their targets. The act of cyberbullying according to Cambridge dictionary is defined as the activity of using internet to harm, frighten or bully another person, especially by sending them unpleasant messages. The effect it has on his victim is the main fact that distinguishes cyberbullying from the normal bullying. It may happen traditional bullying may end in physical damage as well as emotional and psychological damage, as oppose to cyberbullying, where it is all emotional and psychological. It has to be prevented given the consequences of cyberbullying, it urgently needs to get detected as soon as it happens. One of the successful approach that learns from its data and generates a model that automatically classify proper action is Machine learning. It can lead us to detect a pattern among the language of bullies and hence can generate a model to detect cyberbullying actions. Thus, the main contribution of our paper is to propose a supervised machine learning approach for detecting and preventing cyberbullying.

1.1 Data Overview

We performed experiments using large, diverse publicly available datasets for cyberbullying detection in social media. As twitter dataset contains examples of racism and sexism. Where pot labeled as cyberbullying are very less compared to normal neutral posts. Variation Differences in the number of records in the data sets also affect the size of the dictionary, which represents the number of different words found in the data set. We measure the size of the post in terms of the number of words in the post. There are only a few large messages for each data set. We truncate such large posts to the size of a post with a rating of 95 in this dataset. This dataset that we are working upon is taken from Kaggle, it is a twitter dataset having 8818 annotated tweets and he bullying ones having common slurs and terms used pertaining to religious, sexual, gender, and ethnic minorities

1.2 The Approach

We first worked on applying sentiment analysis on Twitter. We worked on a supervised approach to solve this problem. In this approach, we used a dataset containing different segments, namely, id, tweet, label ('0' for positive and '1' for negative). Then we pre-processed the data in different stages. Firstly, we combined the test and the training datasets to create a new data-frame. Then, we removed the user handles (represented after @) as they don't contribute in analyzing the tweet. Further we removed all the punctuations, numbers and special characters in order to create a clean database for implementation. Also, short words don't contribute much in understanding the tweet by the algorithm, so we removed all the words with word length of less than 3 characters. After obtaining a clean dataset, we implemented tokenization of the data because we will apply stemming from the 'NLTK' package. After stemming, we stitch back the tokens to form the meaningful text for the processing. Afterwards, we use Word Cloud for Data visualization of our current analysis of the tweets in the data frame. The tweets with negative comments are collected and can be transmitted to authorities for checking.

2. TECHNOLOGY USED

Natural language processing (NLP) is a study of artificial intelligence that helps machines and computers understand, interpret, and manipulate simple human language. Natural language processing helps developers organize knowledge to perform tasks such as translation, summarization, named entity recognition, relationship extraction, voice recognition, topic segmentation, etc. Natural language processing is a way that computers analyze, understand, and derive meaning

from day to day human language. Languages such as English, Spanish, Hindi, etc. The main components of natural language processing are:
Morphological and lexical analysis
Syntactic analysis
Semantic analysis
Speech integration
Pragmatic analysis

1) Morphological and lexical analysis

Lexical analysis is a vocabulary that includes your words and expressions. Represents the analysis, identification and description of the structure of words. It includes dividing a text into paragraphs, words, and sentences. Individual words are broken down into their components, and symbols that are not words, such as punctuation marks, are separated from words.

2) Semantic analysis

Semantic analysis is a structure created by the parser that assigns meanings. This component of NLP transfers linear sequences of words into simpler structures. Shows how words are associated with each other.

Semantics focuses only on the literal meaning of words, phrases, and sentences. This only abstracts the meaning of diction or the actual meaning of the given context. The structures assigned by the parser always have assigned meaning

For example, "colorless green idea". This would be rejected by Symantec's analysis as colorless here; Green doesn't make any sense.

3) Pragmatic analysis

Pragmatic analysis deals with the general communicative and social content and its effect on interpretation. It means abstracting or deriving meaningful use of language in situations. In this analysis, the main focus is always on what was said, reinterpreted on what is understood.

Pragmatic analysis helps users discover this intended effect by applying a set of rules that characterize cooperative dialogues.

For example, "close the window?" In the example it should be precisely interpreted as a new request instead of an order.

4) Syntax analysis

Words are commonly accepted as the smallest syntax units. Syntax refers to the principles and rules that govern the sentence structure of any individual language.

The syntax focuses on the proper order of words that can affect their meaning. This involves the analysis of the words in a sentence following the grammatical structure of the sentence. Words are transformed into structure to show how words are related to each other.

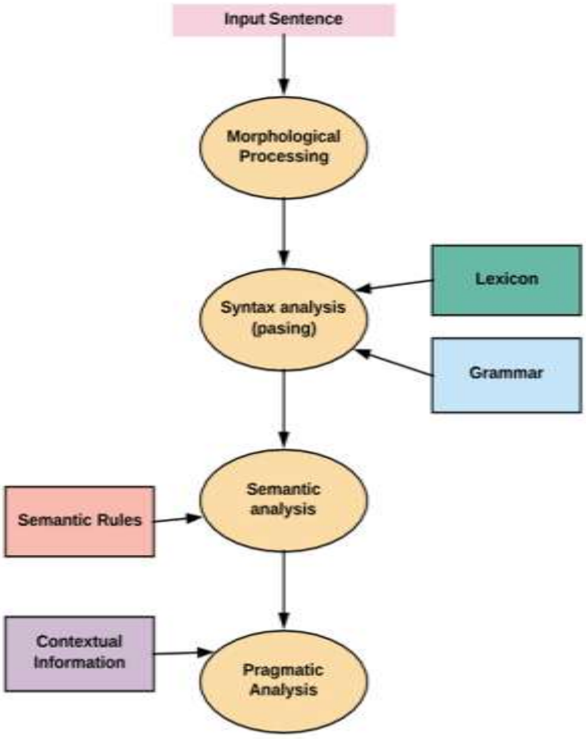

Fig -1: NLP FLOW

3. THE NEXT STEPS

For our current simulation we are using supervised classification of the data. For the next expansion, we are looking for jumping into the unsupervised classification using k-means algorithm. Also, we are going to live track the tweets by integrating the algorithm with the twitter api. After a successful implementation on twitter, we will upgrade the platform for other top social media sites as well. Currently we are working with NLP, but modern-day posts are not limited to texts, but most of the bullying as recorded these days is by the use of images as well. We will also try and build a model for image recognition to further identify and investigate such cybercrime.

4. RESULTS

After preprocessing the dataset, we followed the same steps as were told in section 1.2, we then split the dataset into ratios (0.7,0.3) for train and test. Accuracy, f-score and recall and precision are taken as a performance measure to evaluate the classifiers.
We applied SVM, Random Forest classifier and KNN algorithms as they are among the best performance

classifiers. We achieved highest accuracy while training with SVM and Random forest classifier of 95% whereas KNN gave us the accuracy of 92%.

CLASSIFIER	ACCURACY
SVM	95%
Random Forest	95%
KNN	92%

5. CONCLUSIONS

In this paper, we proposed an approach to detect cyberbullying using machine learning techniques. We evaluated our model on three classifiers SVM, Random forest and KNN and NLTK for feature extraction.

Furthermore, we compared our work with another related work that used the same dataset as ours, finding that our Neural Network outperformed their classifiers in terms of accuracy and f-score. By achieving this accuracy, our work is definitely going to improve cyber-bullying detection to help people to use social media safely.

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. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[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-sne. 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.

[15] 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.

[16] E. Whittaker and R. M. Kowalski. Cyberbullying via social media. Journal of School Violence, 14(1):11–29, 2015.

[17] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In WWW, pages 1391–1399, 2017.

[18] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. Detection of harassment on web 2.0. In The workshop on Content Analysis in the WEB 2.0, WWW, pages 1–7, 2009.

[19] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. In COLING, pages 3485–3495, 2016.