INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Hate Speech on Twitter a Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection

Uppari Kavya
B-Tech Student

Maddula Siva Kesava Reddy
B-Tech Student

Daggumati Sudheer
B-Tech Student

Dama Shanmukha Chowdary
B-Tech Student

K.Srinu
(Assistant professor)

Department of Information Technology
CMR Technical Campus
Kadlakoya (V), Medchal, Hyderabad-501401

Abstract: With the rapid growth of social networks websites, communication between people from different cultural and psychological backgrounds became more direct, resulting in more and more "cyber" conflicts between these people. Consequently, hate speech is used more and more, to the point where it became a serious problem invading these open spaces. This hate speech against their beliefs and religion etc. While most of the online social networks websites forbid the use of hate speech, the size of these networks and websites makes it almost impossible to control all of their content. Therefore, arises the necessity to detect such speech automatically and filter any content that presents hateful language or language inciting to hatred. In this project, we propose an approach to detect hate expressions on Twitter. Our approach is based on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm.

I. INTRODUCTION

Online social networks (OSN) and micro blogging websites are attracting internet users more than any other kind of website. Services such those offered by Twitter, Facebook and Instagram are more and more popular among people from different backgrounds, cultures and interests. Their contents are rapidly growing, constituting a very interesting example of the so-called big data. Big data have been attracting the attention of researcher, who have been interested in the automatic analysis of people's opinions and the structure/distribution of users in the networks, etc. While these websites offer an open space for people to discuss and share thoughts and opinions, their nature and the huge number of posts, comments and messages exchanged makes it almost impossible to control their content. Furthermore, given the different backgrounds, cultures and believes, many people tend to use and aggressive and hateful language when discussing with people who do not share the same backgrounds. King et al. reported that 481

hate crimes with an anti-Islamic motive occurred in the year that following 9/11, 58% of them were perpetrated within two weeks after the event. However, nowadays with the rapid growth of OSN, more convicts' are taking place, following each big event or other. Nevertheless, while the censorship of content remains a controversial topic with people divided into two groups, one supporting it and one opposing it, in OSN, such languages still exists. It is even easier to spread among young people as well as older ones than other "cleaner" speeches. For these reasons, Burlap et al. claimed that collecting and analyzing temporal data allows decision makers to study the escalation of hate crimes following "trigger" events. However, "official" information regarding such events are scarce given that hate crimes are often

ISSN: 2057-5688

unreported to the police. Social networks in this context present a better and more rich, yet less reliable and full of noise, source of information. To overcome this noise and then on-reliability of data, we propose inthis work an efficient way to detect both offensive posts and hate speeches in Twitter. Our approach relies on writing patterns, and unigrams along with sentimental features to perform the detection. The remainder of this paper is structured as follows: in Section 2 we present our motivations and describe some of the related work. In Section 3 we formally define the aim of our work and describe in detail our proposed method for hate speech detection and how features are extracted. Concludes this paper and proposes possible directions for future work.Proposed System:In this project, we propose an approach to detect hate expressions on Twitter. Our approach is based on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm. Our approach relies on writing patterns, and unigrams along with sentimental features to perform the detection.

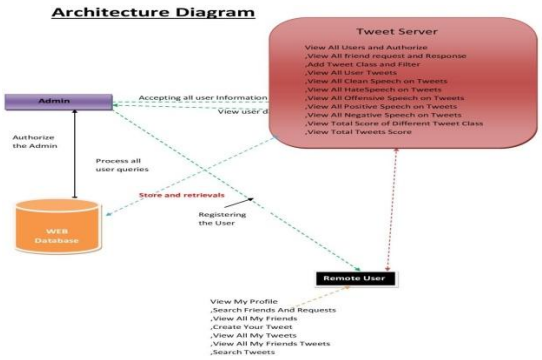
II. LITERATION SURVEY

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate Speech Detection with Comment Embeddings," in Proc. WWW'15 Companion, pp. 29–30, May 2015. The analysis of subjective language on OSN has been deeply studied and applied on different fields varying from sentiment analysis [10] [11] [12] to sarcasm detection [6] [7] or detection of rumors [13] etc. However, relatively fewer works (compared to the aforementioned topics) have been addressed to the hate speech detection. Some of these works targeted sentences in the world wide web such as the work of Warner et al. [5] and Djuric et al. [14]. The first work reached an accuracy of classification equal to 94% with an F1 score equal to 63.75% in the task of binary classification, and the second reached an accuracy equal to 80%. Njagi Dennis Gitari, Z. Zuping, Hanyurwimfura Damien, and Jun Long, "A Lexicon-based Approach for Hate Speech Detection," in , pp., Apr. 2015. Gitari et al. [15] extracted sentences from some major "hate sites" in United States. They annotated each of the sentences into one of three classes: "strongly hateful (SH)", "weakly hateful (WH)", and "non-hateful (NH)". They used semantic features and grammatical patterns features, run the classification on a test set and obtained an F1-score equal to 65.12%. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang, "Abusive Language Detection in Online User Content," in Proc. WWW'16, pp. 145–153, Apr. 2016. Nobata et al. [16] used lexicon features, n-gram features, linguistic features, syntactic features, pretrained features, "word2vec" features and "comment2vec" features to perform the classification task into two classes, and obtained accuracy equal to 90%.Nevertheless, some other works targeted the detection of hateful sentences in Twitter. Kwok et al. [17] targeted the detection of hateful tweets against black people. They used unigram features which gave an accuracy equal to 76% for the task of binary classification. Obviously, the focus on the hate speech toward a specific gender, ethnic group, race or other makes the collected unigrams related to that specific group. Therefore, the built dictionary of unigrams cannot be reused to detect hate speech towards other groups with the same efficiency. Burnap et al. [3] used typed dependencies (i.e., the relation between words) along with bag of words (BoW) features to distinguish hate speech utterances from clean speech ones.

III. IMPLEMENTATION

System architecture

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688



Modules

Admin

In this module, the Admin has to login by using valid user name and password. After login successful he can perform some operations such as View All Users And Authorize, View All friend request and Response, Add Tweet Class and Filter, View All User Tweets, View All Clean Speech on Tweets, View All Hate Speech on Tweets, View All Offensive Speech on Tweets, View All Positive Speech on Tweets, View All Negative Speech on Tweets, View Total Score of Different Tweet Class, View Total Tweets Score.

User

In this module, there are n numbers of users are present. User should register before performing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user can perform some operations like My Profile, Search Friends and Requests, View All My Friends, Create Your Tweet, View All My Tweets, View All My Friends Tweets, Search Tweets.

Searching Users to make friends

In this module, the user searches for users in Same Network and in the Networks and sends friend requests to them. The user can search for users in other Networks to make friends only if they have permission.

Friend Request & Response

In this module, the admin can view all the friend requests and responses. Here all the requests and responses will be displayed with their tags such as Id, requested user photo, requested user name, user name request to, status andtime & date. If the user accepts the request then the status will be changed to accept or else the status will remains as waiting.

IV. RESULT

FIG 8.3. All Positive Speech

FIG 8.4. All Hate

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Speech

FIG 8.5. All Negative Speech

FIG 8.6. All Offensive Speech

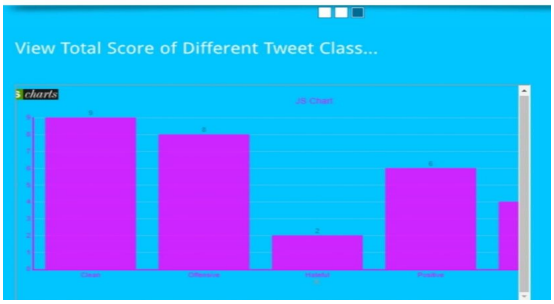


FIG 8.7. Total Score of Different Tweet class

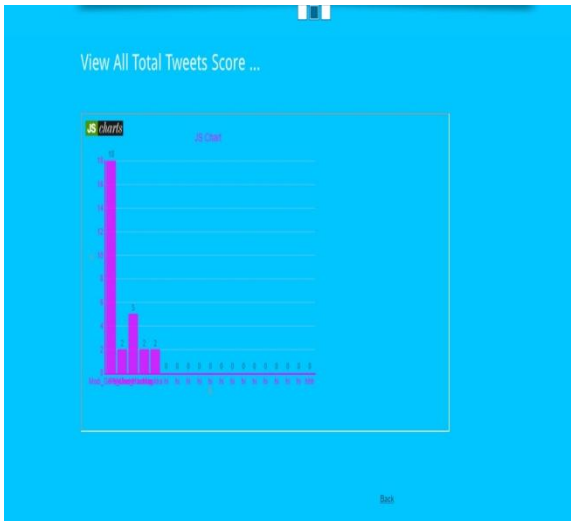


Fig-8.8. Total Tweets Score

V. CONCLUSION:

In this work, we proposed a new method to detect hate speech in Twitter. Our proposed approach automatically detects hate speech patterns and most common unigrams and uses these along with sentimental and semantic features to classify tweets into hateful, offensive and clean. Our proposed approach reaches an accuracy equal to 87.4% for the binary classification of tweets into offensive and no offensive, and an accuracy equal to 78.4% for the ternary classification of tweets into, hateful, offensive and clean. In a future work, we will try to build a richer dictionary of hate speech patterns that can be used, along with a unigram dictionary, to detect hateful and offensive online texts. We will make a quantitive study of the presence of hate speech among the different genders, age groups and regions, etc.

REFERENCES

1. R.D. King and G.M. Sutton, “High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending”,in Criminology pp. 871–894, 2013.
2. PeterJ.Breckheimer,“A Haven for Hate: The Foreign and Domestic Implications

ISSN: 2057-5688

of Protecting Internet Hate Speech Under the First Amendment," in South California Law Review, vol. 75, no. 6, Sep. 2002.
3. P. Burnap, and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statisitcal Modeling for Policy and Decistion Making," in Policy and Internet pp. 223– 242, June 2015.
4. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin, "Offensive Language Detection Using Multi-level Classification," Advances in Artificial Intelligence, vol. 6085, pp. 16–27, Springer, Ottawa, Canada, June 2010.
5. W. Warner and J. Hirschberg "Detecting hate speech on the World Wide Web," in Proc. Second Workshop Language Social Media, pp. 19– 26, June 2012.
6. M.BouaziziandT.Ohtsuki,"Apattern-basedapproachforsarcasm detection on Twitter," IEEE Access, Vol. 4, pp. 5477–5488, 2016.
7. D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcastic sentences in Twitter and Amazon," In Proc.14th Conf. on Computational Natural Language Learning, pp. 107–116, July 2010.
8. M. Bouazizi and T. Ohtsuki, "Sentiment Analysis: from Binary to Multi-Class Classification - A Pattern-Based Approach for MultiClass Sentiment Analysis in Twitter," in Proc. IEEE ICC, pp. 1–6, May 2016.
9. M. Bouazizi and T. Ohtsuki, "Sentiment analysis in Twitter: from classification to quantification of sentiments within tweets," IEEE Globecom, Dec. 2016, to be published.