A Preventive Measure on Hate Speech Detection On Online Social Network using Naïve Bayes

Nupur Khond, Godawari Padwal, Veena Ulgekar,
Tejaswini Parsekar and Sumit Harale

March 15, 2020

A Preventive Measure on Hate Speech Detection On Online Social Network using Naïve Bayes

Nupur Khond
Computer Engineering
Indira college of engineering and management
Pune,Maharashtra
nupurkhond@gmail.com

Godawari Padwal
Computer Engineering
Indira college of engineering and management
Pune,Maharashtra
godawaripadwal22@gmail.com

Veena Ulgekar
Computer Engineering
Indira college of engineering and management
Pune,Maharashtra
veenaulgekar24@gmail.com

Tejaswini Parsekar
Computer Engineering
Indira college of engineering and management
Pune,Maharashtra
tejaswiniparsekar89@gmail.com

Prof. Sumit Harale
Indira college of engineering and management
Pune,Maharashtra
sumit.harale@indiraicem.ac.in

***Abstract*- Online Social network sites are a perfect vicinity for net users to hold in touch, proportion information about their day by day activities and pursuits, publishing and having access to documents, photos and videos[3]. OSN like Facebook, Twitter, Instagram and Google give users the freedom to express their thoughts in text without following traditional language grammar, thereby making it difficult to mine social media for insights. Nevertheless, online social platforms are beset with hateful speech - content that expresses hatred for a person or group of people. Such content can frighten, or may create silence among platform users, and some of it can incite other users to commit violence[2]. Where we have developed a model in which will perform detection on the comments posted on a post based on certain criterions of words distinguishing them whether they are vulgar,offensive or hateful Also a contribution to this will add more functionality to the project[1]. This paper proposes a model which will detect hateful words using Naive Bayes and prevent the hateful comments by hiding them by using hiding mechanism. This work is important because only detecting and identifying hate words is not enough, some actions must be taken against those hateful comments.**

Index: OSN, Naive Bayes,Hateful.

I) INTRODUCTION

Online Social network websites are simply a ideal location for internet users to hold in touch, percentage records about their each day activities and pursuits, publishing and gaining access to documents, photos and videos. OSN like Facebook, Twitter, Instagram and Google give users the liberty to specify their thoughts in text without following traditional language grammar, thereby making it difficult to mine social media for insights.OSN - collectively are some of the most visited websites. Regrettably, OSN are also the right plaza for proliferation of dangerous facts. Cyber-bullying, sexual predation, self-harm practices incitement are a number of the powerful consequences of the dissemination of malicious data on OSN. A lot of those attacks are often carried via a single man or woman, but they can be also controlled by means of companies. The goal of the trolls are often decided on sufferers however, in some situations, the hate can be directed in the direction of extensive groups of individuals, discriminated against for some functions, like race or gender. Such campaigns may also involve a big number of haters which might be self-excited by way of hateful discussions, and such hate would possibly become a physical violence or violent movement. inside the area of research of terrorism and expertise the way terrorist agencies speak with the general public in order to steer human beings in their right to resistance the use of their way of choice, it's far essential a good way to discover the presence of hate speech as a way of implementing thoughts to the public (without delay or in a round-about way), generalizing guilt and radicalizing humans if you want to receive terrorist practices or maybe grow to be individuals of terrorist companies. Moreover, given the one-of-a-kind backgrounds, cultures and beliefs, many people have a tendency to use aggressive and hateful language when discussing with folks that do not now share the identical backgrounds. Namely, 481 hate crimes with an anti-Islamic motive passed off within the 12 months that following 9/11, 58% of them had been perpetrated inside two weeks after the event.Nevertheless, while the censorship of content remains a controversial subject matter with human beings divided into agencies, one assisting it and one opposing it, on OSN, such language nonetheless exists.Hence detecting such comments is not enough but we have to take some action to prevent those hateful comments,to do so we are using this approach for detecting and hiding those hateful comments, we have used Machine Learning Algorithm i.e. Naive Bayes Classifier for detecting the hateful words and hiding mechanism to hide the comments.

II) RESEARCH METHODS

A). Naïve *Bayes*

In machine learning, the Naïve Bayes classifier is a part of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. Naive Bayes is a simple technique for constructing classifier models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a group of algorithms based on common principles. All Naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. But disadvantage of this is even if the fruit satisfies 2 of the three properties but fails the third because of size feature then the detection predicted is also output wrong. A Naïve Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness, and diameter features.

Limitations

1. Hateful speech detection is only on comments.

2 Its detection is only on textual format and not on image.

Disadvantages of existing system

1. Despite the lack of consensus abusive speech.

2. Security weakness.

Mathematical Derivation of Naive Bayes is based on bayes theorem is defined as follows

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Where,

P(A|B) is the posterior probability

P(B) is the predictor prior probability

P(B|A) is the likelihood

P(A) is the class prior probability.

hence we can derive the expression as follows

$$P(A|B) = P(B1|A) \times P(B2|A) \times P(B3|A) \times P(B4|A) \ldots\ldots P(Bn|A) \times P(A)$$

To understand this consider the following example

Naïve Bayes is a classifier. It predicts probability for each class. In hate speech detection consider there are three types of classes:-

A) Vulgar
B) Offensive
C) Bad words

Working of Naïve Bayes:-

1) Training data set will be converted to frequency table.

2) TRAINING DATASET-

Words	Condition
Vulgar	Yes
Offensive	No
Bad	Yes
Vulgar	Yes
Offensive	No
Bad	Yes
Bad	Yes
Bad	No
Offensive	Yes
Vulgar	No
Offensive	Yes

Fig 1.Training Dataset

3) FREQUENCY TABLE

Words	Yes	No
Offensive	2	2
Vulgar	2	1
Bad	3	1
Total	7	4

Fig 2. Tested Data

Naïve Bayes equation is used to predict probability of each class.

The equation is as follows -P(c|x) = P(x|c) * P(c) / P(x).

P(Yes|Vulgar) = P(Vulgar|Yes) * P(Yes) / P(Vulgar) P(Vulgar|Yes) = 2/7 =0.28

P(Yes) = 7/11 =0.63

P(Vulgar) = 3/11 =0.27

Therefore,

P(Yes|Vulgar)=0.28*0.63/0.27=0.65

Similarly, Naïve Bayes will calculate probability for all other remaining classes with respect to yes or no.

OBJECTIVE:

1. To detect hateful words in comments.

2. To reduce the use of offensive words on online social networks.

3. To learn machine learning algorithms.

4. To enhance coding skills in java.

B) Hiding mechanism

The user will login to the system. After the user logs in the user will be able to post images. There will be various comments received on the image, these all comments will be stored into the system's database. After the comments are stored in the database, the comments will be checked if there are any vulgar words present in the comment using Naïve Bayes algorithm. If there are any vulgar words present in the comment then a pop up will raise and the comment will undergo a hiding process and that respective comment will get hidden from the user but will still remain in the database. After checking if there are no vulgar words present in the comment then the respective comment will get posted.

C) LITERATURE SURVEY

[1] *Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection.*

Author:- Hajime Watanabe, Mondher Bouazizi, Tomoaki Ohtsuki.

Description:-With the rapid growth of social networks and microblogging 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. Hate speech refers to the use of aggressive, violent or offensive language, targeting a specific group of people sharing a common property. Their 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. Their experiments on a test set composed of 2 010 tweets show that their approach reaches an accuracy equal to 87.4% on detecting whether a tweet is offensive or not (binary classification), and an accuracy equal to 78.4% on detecting whether a tweet is hateful, offensive or clean (ternary classification).

[2] *Using Naïve Bayes Algorithm in detection of Hate Tweets.*

Author:- Kelvin Kiema Kiilu, George Okeyo, Richard Rimiru, Kennedy Ogada

Description:- Social Media has become a very powerful tool for information exchange as it allows users to not only consume information but also share and discuss various aspects of their interest. Nevertheless, online social platforms are beset with hateful speech - content that expresses hatred for a person or group of people. Such content can frighten, intimidate, or silence platform users, and some of it can incite other users to commit violence. The main goal of the author is to develop a reliable tool for detection of hate tweets. This paper develops an approach for detecting and classifying hateful speech that uses content produced by self-identifying hateful communities from Twitter. Results from experiments showed Naive Bayes classifier achieved significantly better performance than existing methods in hate speech detection algorithms with precision, recall, and accuracy values of 58% , 62% ,and 67.47 %, respectively.

[3] *Detecting Hate Speech within the Terrorist Argument: A Greek Case*

Author:- Ioanna K. Lekea, Panagiotis Karampelas .

Description:- The author presents a methodology for automatically detecting the presence of hate speech within the terrorist argument.

Hate speech can be used by a terrorist group as a means of judging possible targets' guilt and deciding on their punishment, as well as a means of making people to accept acts of terror or even as propaganda for possibly attracting new members.

To decide on how the automatic classification will be performed, they experimented with different text analyzing techniques such as critical discourse and content analysis and based on the preliminary results of these techniques a classification algorithm is proposed that can classify the communiqués in three categories depending on the presence of hate speech. The methodology was tested over the existing dataset with all the communiqués and the corresponding results are discussed.

[4] *Identification of Hate Speech in Social Media.*

Author:- N.D.T. Ruwandika1, A.R. Weerasinghe2

Description:- An exploration of different approaches to detect hate speech in social media is presented by the author. Due to the rapid growth of online content, hate speech has become a common issue which can influence a variety of hate crimes. So, there is a need to find an accurate and efficient technique to detect online hate content and flag them automatically. Then a comparison of both supervised and unsupervised learning techniques with different feature types for the task of hate speech detection was done. From all the supervised and unsupervised models Naïve Bayes classifier with TF-IDF features performed best with an F-score of 0.719.

[5] *Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TF IDF based Approach.*

Author :- Aditya Gaydhani, Vikrant Doma, Shrikant Kendre and Laxmi Bhagwat.

Description:- Toxic online content has become a major issue in today's world due to an exponential increase in the use of the internet by people of different cultures and educational backgrounds. Differentiating hate speech and offensive language is a key challenge in automatic detection of toxic text content. In this the author proposed an approach to automatically classify tweets on Twitter into three classes: hateful, offensive and clean. Using Twitter dataset, they perform experiments considering n-grams as features and passing their term frequency-inverse document frequency (TFIDF) values to multiple machine learning models. They perform comparative analysis of the models considering several values of n in n-grams and TF IDF normalization methods. After tuning the model giving the best results, they achieve 95.6% accuracy upon evaluating it on test data. They also created a module which serves as an intermediate between user and Twitter.

III) SYSTEM DESIGN AND OVERVIEW

A)System Overview

System allows the user to register ,after registration the user logs in to the system using mail id and password. After successfully logging in to the system user can view images, post images, view comments, post comments. Users can comment on others' posts as well and when the user comments on others' post, the system checks the comments using naive bayes algorithm and performs detection and if it is offensive words then the comment is hide using the hiding mechanism. If the comments don't have any hateful words then system posts the comment successfully.

B)Data and Comment Classification

Data is used from different social media accounts where the comments are high . Data is classified as vulgar,bad,offensive. The total data is 1000 comments out of which 650 comments are offensive and 350 are on offensive. Then based on the type of data comments are taken into consideration the data is labelled manually. The dataset is then stored into the database and then the naive bayes algorithm is applied for further detection classification of the words by comparing the data from the dataset and the comments uploaded on the OSN.

C)Getting data:

Data of hateful words is collected which is stored in the database ,is compared with the comments posted on the posts ,based on the hateful words which are stored in the database the hateful comments are detected.Dataset stored in the database consist of only offensive words through which the comparison with the comments is done due to which detecting of hateful comments through our system has become easier

IV) IMPLEMENTATION AND TESTING

User uses this system for connecting to other users world wide,,where the user logs in to the system by registering to the system and then by using the credentials logs in to the system after logging to the system user post the images and also posts comments on other images posted by the other users, when the users comment on the images posted by the other users ,the system applies naive bayes algorithm which calculates the probability of the commets and checks for the vulgar comments and if the probability determined is yes then the system gives an alert for posting the vulgar comments and hence the vulgar comments are hidden from the system,the system also provides the pie chart of the comments commented on the posts.

A) UNIT TESTING

It is a software testing in which individual components of a software are tested.

The aim of unit testing is to test or verify if all the components of the software are working as expected.Unit is the small testable part of the software.In unit testing there are multiple inputs which are provided to the system and single output is produced.Unit testing helps to find software bugs easily and quickly.It helps to reduce number of bugs.In this project we are performing unit testing on different modules such as posting comments,posting images,like comments,friend requests.These all modules are tested separately.

B) INTEGRATION TESTING

In integration testing all the modules of the software are combined together as a group and tested.All the modules are integrated and tested in order or finding the bugs.The aim of integration testing is to check the overall performance of the system.Testing is performed to expose the defects.In this project all components which are mentioned in unit testing are integrated together and tested as a whole system to check if the system works properly.

V)RESULTS

The main objective is to examine the comments by performing detection on them by using naïve bayes algorithm which gives the probability of the comments ,and hiding the hateful comments by using certain hiding mechanism. Existing system was based on detecting hateful comments however we have contributed to the system by hiding the hateful comments. The pie chart below shows the percentage of different activities performed by the user and admin on the system .

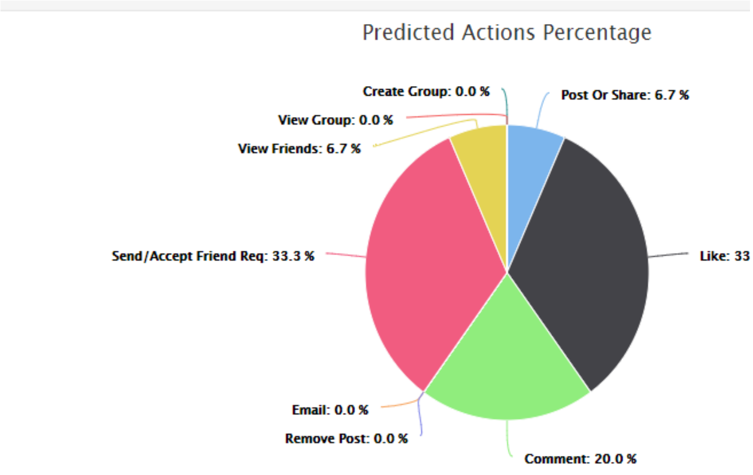

Fig 3. Results

VI) CONCLUSION

In our work, we proposed a new approach to stumble on hate speech in OSN. Social media, as usage of social media has increased to a high extent and hence usage of hateful words has also increased. Our proposed method automatically detects hate speech patterns and most common words and uses these alongside using the above mentioned algorithms to classify comments into hateful, offensive and clean. Computerized detection of abusive language in online social media has in the current years grown to be a key mission. In this contribution we attempted to hide the detected offensive or hateful comment from the user by using certain techniques, such as the tendency to submit abusive messages, as entered to the classifier.

REFERENCES

[1] R. D. King and G. M. Sutton, ``High times for hate crimes: Explaining the temporal clustering of hate-motivated offending," *Criminology*, vol. 51, no. 4, pp. 871_894, 2013.

[2] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy Internet*, vol. 7, no. 2, pp. 223_242, Jun. 2015.

[3] J. P. Breckheimer, "A haven for hate: The foreign and domestic implications of protecting Internet hate speech under the _rst amendment," *South California Law Rev.*, vol. 75, no. 6, p. 1493, Sep. 2002.

[4] W. Warner and J. Hirschberg, ``Detecting hate speech on the World Wide Web," in *Proc. 2nd Workshop Lang. Social Media*, Jun. 2012, pp. 19_26.

[5] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, ``Offensive language detection using multi-level classification," *Advances in Artificial Intelligence*, vol. 6085. Ottawa, ON, Canada: Springer, Jun. 2010, pp. 16_27.

[6] M.Bouazizi and T. O. Ohtsuki, "A pattern-based approach for sarcasm 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. Compute. Natural Language Learning*, Jul. 2010, pp.07_116.

[8] M. Bouazizi and T. Ohtsuki, ``Sentiment analysis in Twitter: From classification to quantification of sentiments within tweets," in *Proc. IEEE GLOBECOM*, Dec. 2016, pp. 1_6.

[9] M. Bouazizi and T. Ohtsuki, ``Sentiment analysis: From binary to multi class classification: A pattern-based approach for multi-class sentiment analysis in Twitter," in *Proc. IEEE ICC*, May 2016, pp. 1_6.