International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

HATE SPEECH DETECTION ON TWEETER

Mr. Supekar Kiran Sampat , Dept of Computer Engineering, PK Technical Campus Chakan, Pune 410501
Mr. Biradar Mahesh Namdev , Dept of Computer Engineering, PK Technical Campus Chakan, Pune 410501
Mr. Yadav Anish Ramashish, Dept of Computer Engineering, PK Technical Campus Chakan, Pune 410501
Mr. Karande Mahesh Namdev , Dept of Computer Engineering, PK Technical Campus Chakan, Pune 410501

ABSTRACT

In recent times, Many countries has been witnessing an insurgence of offensive and hate speech along with racial and ethnic dispositions on Twitter. Popular among the Indian languages used in English. Machine learning has been successfully used to detect Hateful, clean, and offensive speech in several English contexts, the uniqueness of tweets and the similarities among hateful, clean, and offensive speeches require domain-specific English corpus to detect the hateful and offensive speech. And therefore we developed an English corpus from tweets and evaluated different machine learning techniques to detect offensive and hate speech. Character n-gram, syntactic-based features, and multi-tier meta-learning models of support vector machine, logistic regression, random forest, gradient boosting algorithms.

Key Words: Twitter, hate speech, machine learning, Convolutional Neural Network (CNN), sentiment analysis.

1. INTRODUCTION

Still, the debate around the regulation of hateful speeches is ongoing, and still not clear whether the best response to it is through legal measures or other methods such as counter-speech and education. Inattentive of the means of countering it, the harmful evident of hate speech makes its detection critical. Both the volume of content generated online, particularly in social media and the psychological burden of manual moderation supports the need for the automatic detection of offensive and hateful content.

Hate crimes are unfortunately nothing new in society. However, social media, microblogging websites, and other means of online communication have begun playing a big and important role in hate crimes. Anonymously the more online communication including social media enables users to express themselves more freely which is by the way necessary. The ability to freely express oneself is a human right that should be considered, inducing and spreading hate towards another group is an abuse of this liberty.

Many online forums such as Facebook, YouTube, and Twitter consider hate speech harmful and have policies to remove hate speech content. Due to the societal concern and how widespread hate speech is becoming on the Internet, there is strong motivation to study the automatic detection of hate speech. We can reduce the spread of hateful content by automating the detection.

2. MOTIVATION

Hate speech is also a particular form of offensive language where the person using it is basing his opinion either on the racist, or extremist background. Merriam-Webster1 defines hate speech as “speech expressing hatred of a particular group of people.” From a legal perspective, it defines it as a “speech that is intended to insult or intimidate a person because of some trait as race, religion, sexual orientation, nationality, or disability. This being the case, hate speech is considered a worldwide problem that many countries and organizations have been standing up against. With the spread of the internet, and the growth of online social networks, this problem becomes even more serious, since the interactions between people became indirect, and people’s speech tends to be more aggressive when they feel physically safer, not to mention that internet presents for many hate groups sees it as an “unprecedented means of communication of recruiting”. In the context of the internet and social networks, not only does hate speech create tension between groups of people, but its impact can also influence businesses, or start serious real-life conflicts. For those reasons, websites like Facebook, Youtube, and Twitter are not allowed the use of hateful speeches. However, hate speeches are always critical to control and filter all the contents. Therefore, we are trying to automatically detect it.

3. LITURATURE SURVEY:

Using the Twitter dataset, we perform experiments considering n-grams as features and passing their term frequency-inverse document frequency (TFIDF) values to multiple machine learning models. By considering several values of n in n-grams and TFIDF normalization methods of the models. The model gives the best results

by achieving an accuracy of 68.12% upon evaluating it on test data. We also create a module that serves as an intermediate between users and Twitter.
Hate speech in the form of racist and sexist remarks is a common occurrence on social media. Nowadays For that reason, many social media services address such a problem of detecting or identifying hate speech, but the definition of hate speech varies markedly and is largely a manual effort (BBC, 2015; Lomas, 2015). In conjunction with character n-grams for hate speech detection, we analyze the impact of various extra-linguistic features. We also present a dictionary based on the most indicative words in tweets in our data.
``Hate speech detection: Challenges and solutions,"
As online content continues to grow, so does The spread of hate speech is increasing as online content continues to grow through social media. We also identify challenges faced by online automatic approaches in text hate speech detection. Among these difficulties are subtleties in language, differing definitions of what constitutes hate speech, and limitations of data available for training and testing of these systems.
Hate speech refers to the use of aggressive, violent, or offensive language, targeting a specific group of people sharing common property, whether this property is their gender (i.e., sexism), their ethnic group or race (i.e., racism) or they believe and religion.

4. PROPOSED SYSTEM:

4.1 Convolutional Neural Network

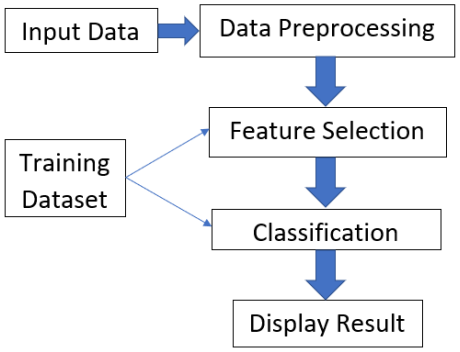
It is assumed that reader knows the concept of Neural Network. When it comes to Machine Learning, Artificial Neural Networks perform really well. Artificial Neural Networks are used in various classification task like image, audio, words. Different types of Neural Networks are used for different purposes, for example for predicting the sequence of words we use Recurrent Neural Networks more precisely an LSTM, similarly for image classification we use Convolution Neural Network. In this blog, we are going to build basic building block for CNN. Before diving into the Convolution Neural Network, let us first revisit some concepts of Neural Network.

5. SYSTEM DESIGN AND OVERVIEW

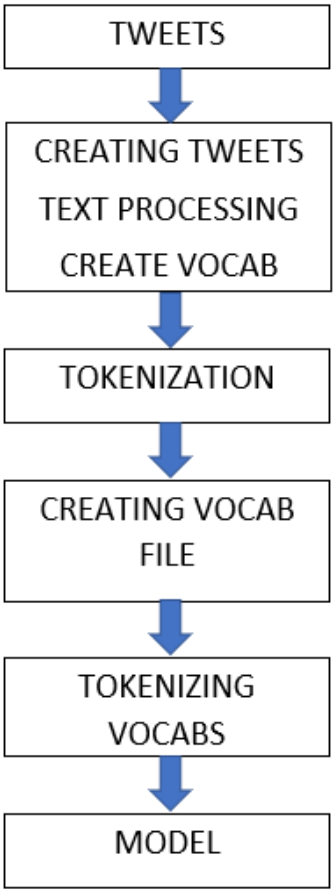
5.1 System Overview

The system will be built in this paper is a system that can classify an Indonesian sentence in an Twitter tweet's section is a hate or not. This system used multinomial logistic regression. TF-IDF is a system for feature extraction from the text. The system implements TF IDF weighting method . The input data used in the form of text in from Twitter tweet's section. The result of this project is classified comment as hate speech or not.

6. SYSTEM ARCHITECTURE



7. DATA PRE-PROCESSING



8.1 ALGORITHM

8.1.1 NAIVE BAYES

Naive Bayes is a simple technique for constructing classifiers: 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 family of algorithms based on a common principle: all naive 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. A naive 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 color, roundness, and diameter features.

8.1.2 LOGISTIC REGRESSION

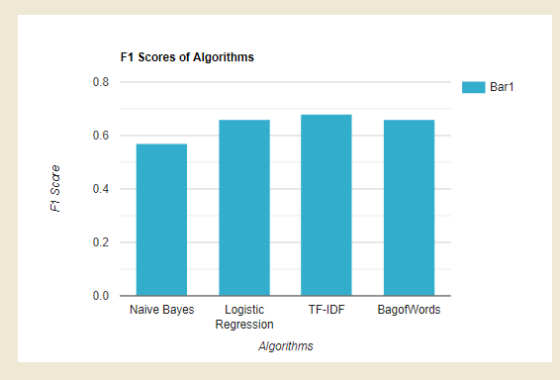
Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

8.1.3 TF-IDF

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents. It has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

8.1.4 CONTINUOUS BAG OF WORDS

The CBOW model tries to understand the context of the words and takes this as input. It then tries to predict words that are contextually accurate. Let us consider an example for understanding this. Consider the sentence: 'It is a pleasant day' and the word 'pleasant' goes as input to the neural network. We are trying to predict the word 'day' here. We will use the one-hot encoding for the input words and measure the error rates with the one-hot encoded target word. Doing this will help us predict the output based on the word with least error.



9. FUTURE WORK

- Explore future work in numerous ways
- Improve accuracy
- Try to build a richer dictionary of hate speech patterns
- Use metadata along with tweets such as number of followers, location, total number of tweets, etc., of a user.
- The dataset can be extended to capture more writing styles, patterns, and topics.

10. CONCLUSIONS

In this work , We proposed a new method to detect hate speech in Twitter. Our proposed approach automatically detects heat speech pattern . We use semantic sentimental feature to classify the tweets into hateful clean Our proposed system reaches an accuracy of 68.12 In future work we will try to build a richer . Dictionary for hate speech patterns we will also they to increase the accuracy of the model by using LSTM algorithm.

11. REFERENCES

[1] S. Dredge. (2014). Twitter Changes: 20 Hits and Misses from the Social Network's History. The Guardian. Accessed: Sep. 10, 2019. Available: https://www.theguardian.com/technology/2014/oct/22/twitterchanges - hits-misses-history

[2] A. Gaydhani, V. Doma, S.Kendre, and L. Bhagwat, ``Detecting hate speech and offensive language on Twitter using machine learning: An N-gram and TFIDF based approach,'' in Proc. IEEE Int. Advance Comput. Conf., Sep. 2018, pp. 1_5.

[3] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, ``SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval),'' in Proc. 13th Int. Workshop Semantic Eval. (SemEval), 2019, pp. 75_86.

[4] Z. Waseem and D. Hovy, ``Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,'' in Proc. NAACL Student Res. Workshop, 2016, pp. 1_6.

[5] S. Macavaney, H.-R.Yao, E.Yang, K. Russell, N. Goharian, and O. Frieder, ``Hate speech detection: Challenges and solutions,'' PLoS ONE, vol. 14, no. 8, Aug. 2019, Art. no. e0221152.
[6] H. Watanabe, M. Bouazizi, and T. Ohtsuki, ``Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection,'' IEEE Access, vol. 6, pp. 13825_13835, 2018.