CYBERBULLYING DETECTION SYSTEM

Oishi Basak
Sahil Arora
Vatsal Agarwal

Mentor: Dr. D. Lakshmi
Affiliations: VIT University Bhopal
SCSE (AIML)

Project Group ID:- 44

ABSTRACT

Many tweets are posted every day and the hate and offensive language occurs more than ever. It is usually very crucial to distinguish different types of abusive language. By classifying tweets, we can know peoples attitude to certain news, celebrities and events in twitter very quickly. For twitter administrators, they can monitor and filter out tweets with extreme language more efficiently based on classifying. For instance, a user can report any tweet with abusive language. These tweets can be classified first before they go to the administrators, which improves the efficiency of the monitor process. For twitter users, there can be a new function to filter these abusive tweets which they may not want to see. A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language. Our project is a classic multi-class classification data mining problem.

INTRODUCTION

India ranked 3rd for cyber-bullying crimes, and this is not even shocking anymore as the number of users are rising day by day and most of them are fake accounts

Cyberbullying is bullying with the use of digital technologies. It can take place on social media, messaging platforms, gaming platforms and mobile phones. It is repeated behavior , aimed at scaring, angering or shaming those who are targeted.

The non-consensual distribution of intimate images involves the sharing of intimate images, often of a former partner, with third parties (either via the Internet or otherwise) without the consent of the person depicted in the image. Often the motivation is to take revenge against their former partner. Its effect is a violation of the former partner's privacy in relation to images, the distribution of which is likely to be embarrassing, humiliating, harassing, or degrading to that person.

MODULES AND METHODS

Web Scraping -
It extracts content and data from a website. Unlike screen scraping, which only copies pixels displayed onscreen, web scraping extracts underlying HTML codes, and with it, data stored in a database

Data Collection –
Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Data Wrangling –
Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. It is also known as data cleaning.

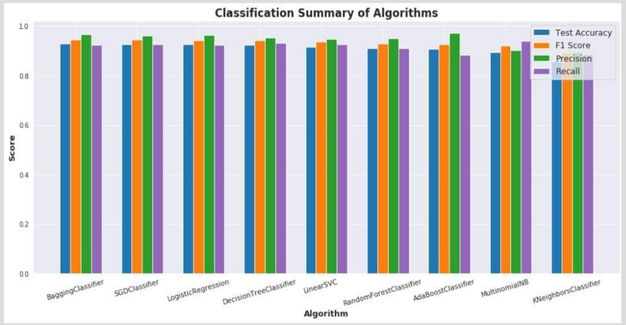
Data Normalization -
Data normalization is a process in which data attributes within a data model are organized to increase the cohesion of entity types

Feature Selection -
Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

Model Building -
Model building is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables.

Deployment-
The concept of deployment in data science refers to the application of a model for prediction using new data.

RESULTS



Best Accuracy: 0.928 - BaggingClassifier
Best F1 Score: 0.944 - BaggingClassifier
Best Precision: 0.973 - AdaBoostClassifier
Best Recall: 0.940 - MultinomialNB

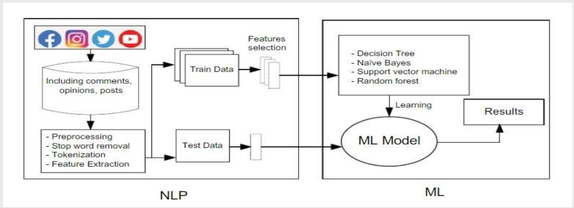


Figure 1: Real Time Usage

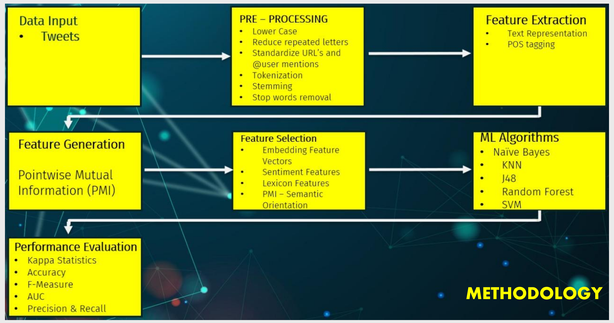


Figure 2: Methodology

DISCUSSION

•This model scraps Twitter so as to scan for tweets.
•After web scraping (with the help of Twint and Tweepy modules in Python and Beautiful Soap and Selenium and Twitter API) we make our .csv file.
•After the .csv file is prepared, we need to preprocess the data and clean the data and distinguish and convert the continuous data into categorical data.
•Then we will divide our dataset in testing and training set for training and testing of our model.
•Then we need to apply Machine Learning Models (such as Support Vector Machines, Logistic Regression, Vectorization, Decision Tree Classifiers, Neural Networks) to detect the presence of hate tweets.
•After the detection, we need to analyze the result.
•We will make a Google extension so that it works on every website in Google. Our extension directly connects the Twitter website to our Machine Learning model and gives the output.
•If a hate tweet is found, our model flashes a warning to the user.
•If the Twitter website has no such hate tweets, it launches smoothly.

CONCLUSIONS

Our model presents a cyberbullying detection technique using a combination of four machine learning architectures (i.e., SGD, Logistic Regression, Linear SVC, Decision Tree and Random Forest Classifier models). The proposed method is evaluated using real-time dataset achieved by web scraping of the Twitter API to have enough samples to train our model. The experimental results show the significance of this method in classifying short messages (e.g., tweets). The proposed method achieved good results compared to the state-of-the-art methods on the dataset, achieving an accuracy of approximately 88% when the dataset was split into 75% training and 25% testing.

REFERENCES

- Neural Models for Offensive Language Detection
  https://arxiv.org/pdf/2106.14609.pdf
- Automated Hate Speech and Offensive Language Detection
  https://www.researchgate.net/publication/314942659_Automated_Hate_Speech_Detection_and_the_Problem_of_Offensive_Language
- A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media
  https://www.researchgate.net/publication/355845224_A_Multichannel_Deep_Learning_Framework_for_Cyberbullying_Detection_on_Social_Media
- Collaborative Detection of Cyberbullying Behavior in Twitter Data Amrita Mangaonkar, Allenoush Hayrapetian, Rajeev Raje
  (PDF) Collaborative detection of cyberbullying behavior in Twitter data (researchgate.net)