TOXIC COMMENT CLASSIFICATION FOR SOCIAL MEDIA USING IBM SERVICES

A MINOR PROJECT REPORT

Submitted to

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY, HYDERABAD

In partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

ELUGAM JYOSHNA	**18UK1AO5D1**
BAREEDULA VIKASINI	**18UK1A05C5**
THATIKONDA GOWTHAM	**18UK1A05D3**
S CHANDANA	**18UK1A05A5**

under the esteemed guidance of

Dr.Rajendar Reddy

(professor)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VAAGDEVI ENGINEERING COLLEGE

BOLLIKUNTA, WARANGAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VAAGDEVI ENGINEERING COLLEGE

(Affiliated to JNTU Hyderabad & Approved by AICTE, New Delhi)

Bollikunta , Warangal – 506005

2018-2022

CERTIFICATE

This is to certify that the Major Project Report entitled "TOXIC COMMENT CLASSIFICATION FOR SOCIAL MEDIA USING IBM SERVICES" is being submitted by E.JYOSHNA (H.NO:18UK1A05D1), B.VIKASINI(H.NO:18UK1A05C5),T.GOWTHAM(H.NO:18UK1A05D3),S.CHANDANA(H.NO:18UK1A05A5) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering to Jawaharlal Nehru Technological University Hyderabad during the academic year 2021-22, is a record of work carried out by them under the guidance and supervision.

Project guide	Head Of Th e Department
Dr.Rajendar Reddy	Dr.R.Naveen Kumar
(professor)	(profess or)

ACKNOWLEDGEMENT

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dr.P.PRASAD RAO, Principal, Vaagdevi Engineering College for making us available all the required assistance and for his support and inspiration to carry out this major project in the institute.

We extend our heartfelt thanks to Dr.R.NAVEEN KUMAR, Head of the Department of CSE, Vaagdevi Engineering College for providing us necessary infrastructure and there by giving us freedom to carry out the major project.

We express heartfelt thanks to Mr.Ch.Jayaprakash, Program Manager, SmartBridge Educational Services Private Limited, for their constant supervision as well as for providing necessary information regarding the major project and for their support in completing the major project ,mini project and internship.

We express heartfelt thanks to the guide, Dr.Rajendar Reddy , professor, Department of CSE for his constant support and giving necessary guidance for completion of this minor project.

Finally, we express our sincere thanks and gratitude to my family members, friends for their encouragement and outpouring their knowledge and experience throughout the thesis.

E.JYOSHNA (18UK1A05D1)

B.VIKASINI (18UK1A05C5)

T.GOWTHAM (18UK1A05D3)

S.CHANDANA (18UK1A05A5)

ABSTRACT

Nowadays, the flow of data over the internet has grown dramatically, especially with the appearance of social networking sites. Social networks sometimes become a place for threats, insults, and other components of cyberbullying. A huge number of people are involved in online social networks.

Toxic comments are textual comments with threats, insults, obscene, racism, etc. In recent years there have been many cases in which authorities have arrested some users of social sites because of the negative (abusive) content of their personal pages. Hence, the protection of network users from anti-social behavior is an important activity. One of the major tasks of such activity is automated detecting the toxic comments. Bag of words statistics and bag of symbols statistics are the typical source information for the toxic comments detection. Usually, the following statistics-based features are used: length of the comment, number of capital letters, number of exclamation marks, number of question marks, number of spelling errors, number of tokens with non-alphabet symbols, number of abusive, aggressive, and threatening words in the comment, etc. A neural network model is used to classify the comments.

1.INTRODUCTION

1.1 OVERVIEW

Over the years, social media and social networking use have been increasing exponentially due to an upsurge in the use of internet. Flood of information arises from online conversation in a daily basis as a people are able to discuss, express themselves and air their opinion via this platforms. While the situation is highly productive an could contribute significantly to the quality of human life, it could also be destructive and enormously dangerously. While discussion or a conversation is opened, it is quite obvious that debates may arises due to differences in opinion. But often these debates take a dirty side and may result in fight over the social media during which offensive language termed as toxic comments may be used from one side. These toxic comments may be threatening, obscene, insulting or identity-based hatred. So, these clearly pose the threat of abuse harassment online. consequently, some people stop giving their opinions or give up seeking different opinion which result in unhealthy and unfair discussion. As a result, different platforms and communities find it very difficult to facilitate fair conversation and are often forced to either limit user comments or get dissolved by shutting down user comments completely. This study focuses on building a multi-headed model to detect different types of toxicity like threats, obscenity, insults, and identity based hate. Detecting and controlling verbal abuse in an automated fashion is inherently a natural language processing task. Natural language processing, (NLP), is a branch of artificial intelligence that deals with the interactions between computers and human using natural language.

1.2 PURPOSE

Toxic comments classification on online channels is conventionally carried out either by moderators or with the help of text classification tools. With resent advances in deep learning (DL) techniques, researches and exploring if DL can be used for comment classification task. Text classification is classic topic for natural language processing and an essential component in many applications, such as web searching, information filtering, topic categorisation and sentiment analysis. Text transformation is the very first step in any form of text classification. The online comments are generally in non-standard English and contain lots of spelling mistakes partly because of typos (resulting from small screens of mobile devices) but more importantly because of the deliberate attempt to write the abusive comments in creative ways to dodge the automatic filters.

The contributions this paper as follows:

1. This research work will collect and differentiate toxic classified comments from non-toxic comments.

2. This paper will develop a multi-headed model to detect different type

2.LITERATURE SURVEY

2.1 Existing Problem

Aggression by text is a complex phenomenon, and different knowledge field try to study and tackle this problem. This analysis of related work focuses on a computer science perspective of aggression identification, a recent emerging area.

Currently, the scientific study of automatic identification of aggressively text using information technology techniques, is increasing. In the study, several related literature are used to express different types of aggression. Some of those are hate. Cyber bullying, abusive language, toxicity, flaming, extremism, radicalization, and hate speech. Despite the differences between those concepts, previous research can give us insight into methods to approach the problem of identifying aggressive interactions. Attention is given to the automatic detection of hate speech. Provides a short, comprehensive, structured and critical overview of the field of automatic hate speech detection in natural language processing. This research found a few dedication works that address the effect of incorporating different text transformation on the model accuracy for sentiment classification. The impact of transformation on text classification by taking into account four transformation and their all possible combinational on news an email domain to observe the classification accuracy.

2.2 Proposed Solution

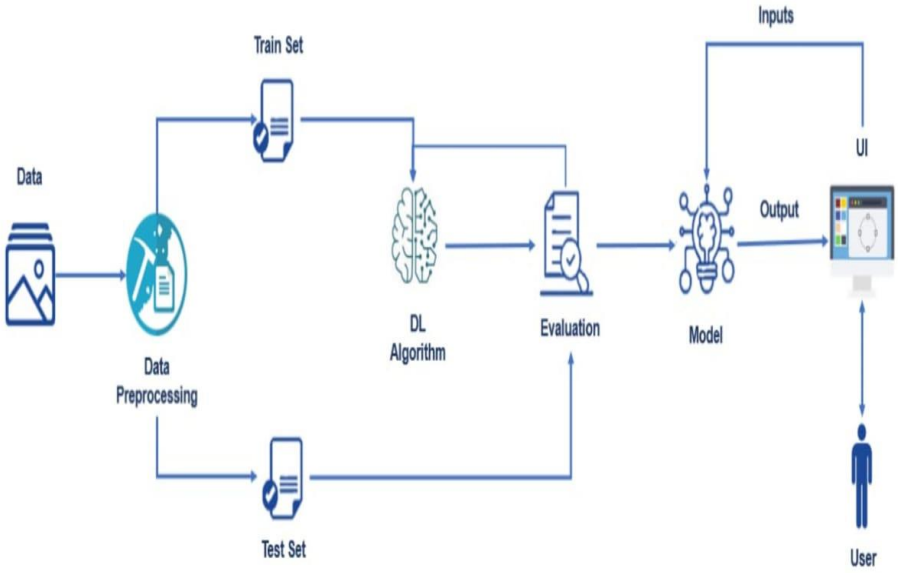
Logistic regression

The logistic regression (LR) algorithm is used for supervised learning, and widely used for binary classification tasks. It is a branch of natural language processing (NLP), which is generally thought of as part of artificial intelligence (AL). LR permits obtaining insights about the model, such as observed coefficients. Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a dataset. Logistic regression has become an important tool in the discipline of machine learning. The approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data.

As more relevant data comes in, the algorithm should get better at predicting classification within data sets. Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specially predefined buckets during the extract, transform, load (ETL) process in order to stage the information for analysis. A logistic regression models predicts a dependent data variable by analysing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college. The resulting analytical model can take into consideration multiple input criteria. In the case of college acceptance, the model could consider factors such as the student grade point average, SAT score and number of extracurricular activities. Based on historical data about earlier outcomes involving in the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

3.THEORITICAL ANALYSIS

3.1 BLOCK DIAGRAM



3.2 HARDWARE/SOFTWARE DESIGNING

1.Importing Dataset

2.Evaluating any null values

3.Training and Testing Dataset

2.model building:

1.Import model building libraries

2.Initialising the model

3.Loading preprocessing data

4.configure learning process

5.Train and Test the Model

3.Application Building:

1.Create HTML file

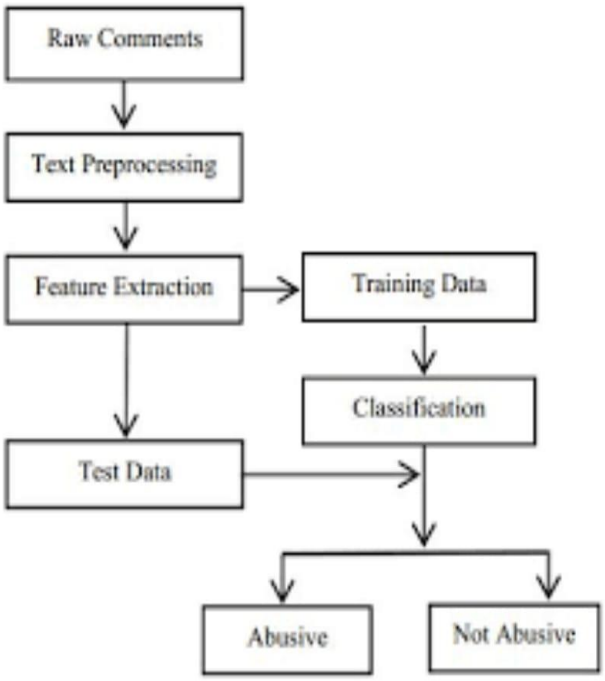
2.Build Python Code

4.EXPERIMENTAL INVESTIGATIONS

The experimental analyses shown that choosing appropriate combination may result in significant improvement on classification accuracy. NOBATA& Tetreault used normalization of numbers, replacing very long unknow words and repeated punctuations with the same token. Explained the role of transformation in sentiment analyses and demonstrated with the help of SVM on movie review database that the accuracies improve significantly with the appropriate transformation and feature selection. They used transformation methods such as white space removal, expanding abbreviation, stemming, stop words removal and negation handling. Other works focus more on modelling as compared to transformation. In the study, Bojanowski et al. used five transformations namely URLs features reservations, negation transformation, repeated letter normalization, stemming and lemmatization on twitter data and applied liner classifier available in WEKA machine learning tool. They found the accuracy of the classification increases when URLs features reservations, negation transformation and repeated letters normalization are employed while decreases when stemming and lemmatization are applied. The investigation the effect of transformation on five different twitter datasets in order to perform sentiment classification and found that removal of URLs,

The removal of stop words and the removal of numbers have minimal effect on accuracy whereas replacing negation and expanding acronyms can improve the accuracy. Most of the expanding regarding application of the transformation has been around the sentimental classification on twitter data which is length restricted.

5.FLOWCHART



Data Collection

- collect the dataset or create the dataset

Data Pre-Processing

- import the libraries
- importing the dataset
- checking for null values
- data visualisation
- taking care of missing data
- splitting data into train and test

Model Building

- Training and testing the model
- Evaluation of Model

Application Building

- Create an HTML file
- Build a Python Code

6.RESULT

Here we used logistic regression to work out on a predicting and analysing toxic comments classification for social media .result mainly tells about,whether the comment given is a toxic comment or not.

7.ADVANTAGES BASED ON PROPOSED SOLUTION

1. In the current century, social media has been created many job opportunities and, at the same time, it has become as unique place for people to freely express their opinions.

2. Comments provide a platform to directly communicate with your customers to show them that u care about their needs in case of business

3. The advantage of comments is the medium that connects blog owners or writers and the users or readers.

4. Commenting has become a new form of written communication as it allows people to avoid being interrupted and express their ideas. People write many comments on social media (on private accounts and beyond them), and it can help brands a lot. Tracking social activity can help you better to serve customers.

5. Comment dialog gives the writer all sorts of springboards for further blog posts. You know that you are writing what your audience is interested in, and that is something every writer hopes to do. Good conversation will have other benefits too.

DISADVANTAGES

1. Social media is toxic because it is additive by design. Also, it is often called toxic because platforms spread disinformation and they encourage their users to engage with these fake news more than with fact checked 'real' news

2. There are some groups that are taking advantage of this framework and misuse this freedom to implement the toxic mindset (insulting, verbal sexual harassment, threads, obscene, etc).

3. The social media comments, mainly which are toxic, based on harassment and body shaming, makes humans hurt, depressed, and kind of health issues.

4. The impact of toxic comments is much more catastrophic then we think. It not only hurts one's self-esteem or deters people from having meaningful discussion, but also provokes people to such sinister.

5. The main disadvantage of comments in social media platform, which makes a fake news viral in a short period of time.

6. It also turns out that exposure to online negativity makes our own thinking negative-reading uncivil comments can immediately increases readers own hostile cognitions. To sum up, we are all subject to social influence online. Reading other people's opinions can influence our own perceptions, thinking and even behaviour.

8.APPLICATIONS

9.CONCLUSION

Communication is one of the basic necessities of everyone's life. People need to talk and interact with one another to express what they think. Over the years, social media and social networking have been increasing exponentially due an upsurge (rise) in the use of the internet. Flood of information arises from online conversation on daily basis, as people are able to discuss, express them selves and express their opinions via these platforms. And while the situation is highly productive and could contribute significantly to the quality of human life, it could also be destructive and enormously dangerous. The responsibilities lies on the social media administration, are the host organisation to control and monitor these comments. This research work focuses on developing a model that would automatically classify a comment as either toxic or non-toxic using logistic regression.

Therefore, this study aim to develop a multi-headed model to detect different types of toxicity like threats, obscenity, insults, and identity-based hate. By collecting and pre processing toxicity classified comments from training and testing

using term frequency inverse document frequency algorithm, developing regression to train and test the dataset.

10.FUTURE SCOPE

The future work contribution of jigsaw is to develop and illustrate a method that combines crowd sourcing and machine learning to analyse personal attacks. This also discusses text mining and, also processing of text carried out using the term frequency-inverse document frequency (TF-IDF) technique. The evaluation of the model is done using confusion metrics.

SCREEN SNAPS

11.BIBLOGRAPHY

[1]. Deng, A., Yu. D., (2014). Deep Learning Applications.

http://research.microsoft.com/pubs/209355/DeepLearn NowPublishing-Vol7-SIG-39.pdf

[2]. Yoshua, B., (2009). Learning Deep Architectures for Foundations and Trends in Machine

[3]. Nobata, C., Tetreault, J., Thomas, A. Mehdad, Y., (2016). Abusive Language Detection in International Conference on

[4]. Aggarwal,C., Zhai, C., (2012). A Algorithms. In Mining Text Data.

[5]. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., (2017). Enriching Word Vectors with Subword Information, TACL, vol

APPENDIX

A. Source code

B. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge