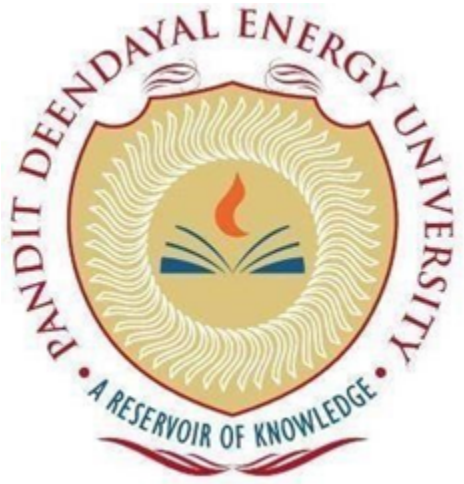
Pandit Deendayal Energy University

School of Technology



Course: Data Mining Lab B.Tech

(Computer Science and Engineering)

Semester 5

"Detection of Possible Toxic Messages on Twitter"

Submitted to:

Dr. Rajeev Gupta

Submitted by:

Nilay Patel - 20BCP005

Shruti Jethloja - 20BCP013

Jai Modi - 20BCP014

Parth Vekaria - 20BCP009

ABSTRACT

The increasing use of social-media platforms in the modern world has increased connectivity from all the different ends of the world. This increase in the accessibility of social media has allowed people to convey their notions to the world. This platform, now-a-days has also become a place where people abuse their anonymity to pass toxic and abusive comments on other people, community, ideas, etc. This paper identifies toxic comments and hate speech as an obstacle in the way of a safe online experience. This paper aims to use modern machine learning methods to identify hate speech and categorize it into categories for easier recognition of hate speech. The study is oriented towards Twitter tweets, as currently, most cases of hate speech transaction has occurred on that platform.

The dataset used in the training purposes is received from a public platform. Multiple preprocessing methods are applied on the set to clean the data and use of Stratified TFIDF vectorizer for feature engineering purposes. The machine learning model used for this purpose is Logistic Regression, a classification model used to classify tweets into categories

1. Toxic
2. Severe Toxic
3. Obscene
4. Threat
5. Insult
6. Identity_Hate
7. Normal(base)

The model, on evaluation, received a CV score of 0.979, which, for text analysis purposes. A few drawbacks in the model are also present. This model is prone to fall for type 1 error.

In future, this model can be trained on several more cases, and implemented as an automated system for social media platforms to generate an extensive report on whether a comment or tweet is toxic in nature or not.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations	Definition
TFIDF	Term Frequency Inverse Document Frequency
LR	Logistic Regression
SVM	Support Vector Machines
CNN	Convolutional Neural Networks
NB	Naive Bayes
KNN	K-Nearest Neighbors
adaBoost	Adaptive Boost
MLP	Multilayer Perception
ML	Machine Learning
NaN	Not a Number
CV	Cross Validation
ROC_AUC	Area under Receiver Operating Characteristic Curve

1 INTRODUCTION

The revolutionary state of the introduction of Web 2.0 had brought about a lot of changes in the world. It has changed our perspective in terms of commerce, communication, and collaboration. This revolution has led to an exponential increase in the communication of thoughts and made the internet the basis of all major communications, be it personal or global. As all coins have two sides to them, so does the revolutionary Web 2.0. Web 2.0 has allowed people to communicate from different corners of the world and convey their notions, where some people use this to enlighten other people, and some use this to pass toxic messages and use anonymity to encourage the flow of haterade in society. The increase in cyberbullying has led people to fear sharing their thoughts online. Currently, Twitter is a platform where such examples are seen on a daily basis regardless of the topic of conversation, hence this study is oriented toward the tweets on Twitter. This study aims to collect real-time tweets and use them from the trained model to categorize those tweets into several categories with increasing severity of toxicity and use this as a report to ban or cancel the senders to decontaminate the fog of hateful internet. The categorization of the tweets shall take place as follows,

8. Toxic
9. Severe Toxic
10. Obscene
11. Threat
12. Insult
13. Identity_Hate
14. Normal(base)

Some studies in the past have already tried to tackle this issue using classifiers, but we, from this study plan to even further those studies by classifying texts into multiple classes rather than fixing them to a particular class, as seen in previous studies. Some studies in the past have tried to figure out the efficient method of application for the detection of hate speed on social media, and we, in this study will be using their path. The feature engineering will be primarily focused on the Stratified TFIDF vectorizer and the machine learning approach used is Logistic Regression.

2 LITERATURE SURVEY

2.1 Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach

Toxic messages are a menace to society, it is reaching heights in the modern world due to exponential increase in the user of the internet. This paper analyzes the key challenge to be the automatic detection of toxic or non toxic tweets, and categorizes the tweets into 3 categories, namely, hateful, offensive, and clean. This study has only classified tweets into 3 categories considering the normal category. In this study, one tweet belongs to only one category of toxic behavior, namely, hateful or offensive. This study has used TFIDF(Term Frequency Inverse Document Frequency) normalization method to convert the tweets into data and used multiple machine learning algorithms and comparative analysis of the models to provide prediction. For the comparative analysis, Logistic Regression, Naive Bayes and Support Vector Machines as classifier models are used, out of which Logistic Regression performs better for n-gram and TFIDF features after tuning the hyperparameters.

Confusion matrix for the evaluated test data on the final logistic regression model

Class	Classified as		
	Hateful	Offensive	Clean
Hateful	0.965	0.021	0.014
Offensive	0.048	0.926	0.026
Clean	0.010	0.013	0.977

2.2 Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites

Human trafficking is a global problem that robs millions of victims of their dignity. The paper focuses on solving a problem of human trafficking by classifying Urls and image based on the gender and age group, to detect the photographs of the children under the age of 14.The paper

focuses on using SVM and Convolutional Neural Network (CNN) along with haar filters, taking into account human torso information and face detection, for classification of gender and age. Data was cleaned by removing noise and misspelled words, and then classified whether suspicious or not by filtering data with special use of hashtags. Spanish data is collected by executing a search request with the following hashtags: #escort, #prepago, #joven (young) , #Dulce(sweet) , #Fresca(new) , #nueva (new), #lolita, and #flaquita. Hashtags were chosen as indicators of underage criteria.

Hashtag	Number of occurrences
#escort	45604
#prepago	15890
#joven	3456
#dulce	1256
#fresca	1456
#nueva	5743
#flaquita	6580
#lolita	867
#penguin	23980
#caldodepollo	45990
#cp	34562

Image classification is performed using two categories :

1. Image classification with face
2. Image classification using upper body feature

For both the classification SVM and CNN model is used. Following are the results of both the classification

- SVM Results

  Confusion matrix for gender (face)

	Man	Woman
Man	160	20
Woman	40	100

  Confusion matrix for gender (upper body)

	Man	Woman14
Man14	220	40
Woman	50	180

Confusion matrix for age group classification (face)

	Over14	Under14
Over14	160	20
Under14	40	100

Confusion matrix for age classification (upper body)

	Over14	Under14
Over14	110	30
Under14	20	120

- CNN Results

CNN performance (face)

FACE	*Accuracy*	*Mean Square Error*
Gender	98,5%	1,2%
Age Group	97,3%	2,8%

CNN performance (upper body)

UPPER BODY	*Accuracy*	*Mean Square Error*
Gender	64,2%	35,8%
Age Group	51,4%	48,6%

As a result, the paper shows higher performance of SVM along with torso features than CNN. Hence SVM model was widely accepted with accuracy of more than 80% for both the classifications.

2.3 Automatic Hate Speech Detection using Machine Learning: A Comparative Study

This study understands the exponential increase in the use of online platforms as a medium of communication. The scholars also realize the backlash provided by such a platform. This study was conducted to fill the gap between an efficient and reliable feature engineering technique and a related machine learning algorithm. The paper focuses on using a publicly available dataset and applying 3 feature engineering techniques and 8 machine learning algorithms and comparing their relative results to name the best method to automatically detect hate speech on social media platforms. The paper has classified the text features into 3 categories, namely,

1. Hate Speech
2. Offensive but not Hate Speech
3. Neither Hate Speech or Offensive(Base class)

The dataset used is as follows

	Class	**Total Instances**	**Training instances**	**Testing instances**
0	Hate Speech	2399	1909	490
1	Not offensive	7274	5815	1459
2	Offensive but not Hate Speech	4836	3883	953
	Total	**14509**	**1607**	**2902**

The text preprocessing methods used in this paper includes converting the entire text into lowercase, as well as removing URL, usernames, white spaces, hashtags, punctuations and stop words. After such, they have applied tokenization and stemming for each preprocessed data.
As suggested before, the feature engineering techniques used are as follows,

1. TFIDF Vectorizer
2. N-gram
3. Doc2Vec

The machine learning algorithms covered by the study includes,

1. Naive Bayes
2. SVM
3. KNN
4. Decision Tree
5. Random Forest

6. AdaBoost
7. Multilayer Perceptron
8. Logistic Regression

According to the conclusion reached by the study, bigram feature engineering method mixed with SVM provides the most precise result with an accuracy of 79%. This study holds as a baseline for future researches to improve upon the accuracy of the Machine Learning model and provide better results

3 PROPOSED WORK

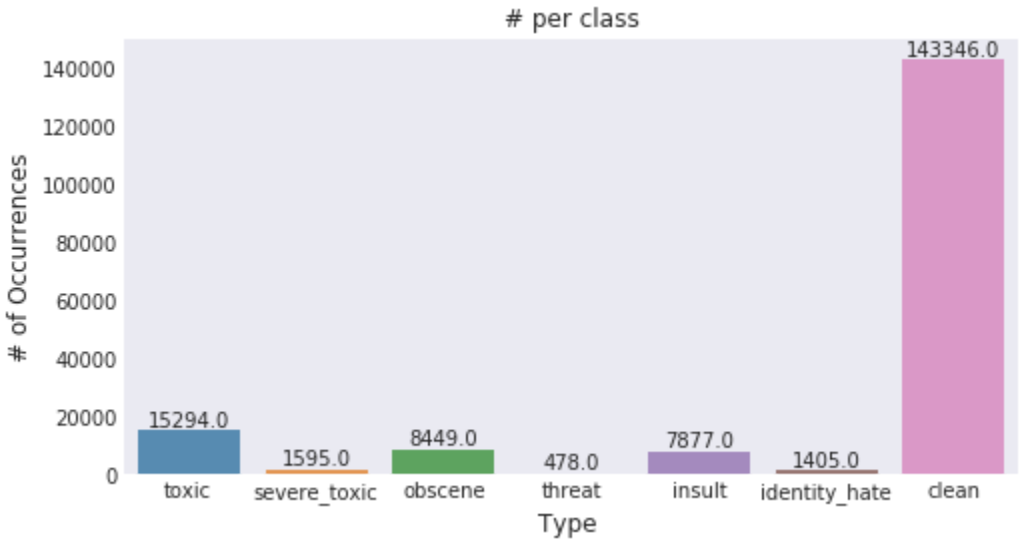
This section elucidates upon the proposed method, or the approach used in this study in order to classify tweets into categories of toxicity. The research methodology is contained in six key steps namely, data collection, data preprocessing, feature engineering, data splitting, classification, and model construction Each of the steps is discussed in detail in the subsequent sections.

3.1 DATA COLLECTION

In this study, the dataset is collected from publically available hate speech dataset. This dataset was pre compiled into multiple labels, where each text can be a part of several categories. Toxic, Severe_toxic, Obscene, Threat, Insult, Identity_hate, Normal. The dataset consists of 312735 instances, with 27493 instances of toxic tweets(belonging to either one of the categories) and the rest normal tweets. Hence the dataset was imbalanced.

3.2 TEXT PREPROCESSING

In this part of the timeline, all the NaN values from the dataset are simply replaced with blank text, as manually assessing each of the NaN input tweets would be more tedious.



The above figure displays the end result of the number of tweets belonging to a particular class after the preprocessing phase.

3.3 FEATURE ENGINEERING

The ML algorithm cannot understand the classification rules from the raw text. These algorithms need numerical features to understand classification rules. Hence, in text classification one of the key steps is feature engineering. This step is used for extracting the key features from raw text and representing the extracted features in numerical form. In this study, we have taken use of the Stratified TFIDF vectorizer.

In order to continue with the multi-label classification, we tried to figure out if there is any correlation between two classes, i.e. if one occurs, is it possible that it explains the occurrence or non-occurrence of another class.

3.4 DATA SPLITTING

Using the normal test train split, in this case might be useless in the dataset collected, as the dataset is imbalanced, for the same reasons Stratified K-Folds Cross Validation is used in order to preserve the percentage of multi label data and train the model to generate a precise model.

3.5 MACHINE LEARNING MODEL

Studies suggest no single classifier can be accurate enough, and have promoted a use of both SVM and Logistic Regression in a hand-in-hand fashion(boosted), but, In this study, our proposed algorithm is LR and as the target was multi-label classification, therefore solver method “sag” was used in order to support multi label classification

4 IMPLEMENTATION DETAIL

As mentioned in chapter 3, the public data received is cleaned in the preprocessing phase, preparing it for the feature engineering phase, here the use of TFIDF vectorizer, where the bag of words is generated and also used this time to decipher any correlation between the classes. Further, the dataset is split into a 67, 33 percent fashion, to be marked as train and test datasets respectively.
Once the human-centered part of cleaning the dataset is over, the machine takes over using the logistic regression model. Now the trained model is evaluated over CV Score (Scoring method is ROC_AUC). This model can now be used to identify hate texts when given input.

As mentioned in chapter 3, the public dataset received is cleaned by filling NaN values with blank comments in the preprocessing phase, preparing it for the feature engineering phase, where the use of Stratified TFIDF vectorizer is used in order to preserve the percentage of all labels in the training dataset. This method was implemented as the received dataset was imbalanced and using a basic test-train split on such a dataset might cause overfitting to one class and underfitting to the other. The model then used for machine learning was LR, as preached by multiple studies in the past. The model was then evaluated on CV score, which uses multiple folds of iteration and returns the average result of all the iterations.

5 RESULT ANALYSIS

The model generated in the process received a CV score of 0.979, which for such text analysis based applications is good. There are places of improvement, such as its prone to type 1 errors. The model is further used to figure out the words with the most frequent use in different classes as shown in the figure below.

```
CV score for class toxic is 0.9692181203716593
CV score for class severe_toxic is 0.9875919864808708
CV score for class obscene is 0.9838682408771228
CV score for class threat is 0.9833768152528718
CV score for class insult is 0.9774237433506198
CV score for class identity_hate is 0.9739428843120509
Total CV score is 0.9792369651075327
```

6 CONCLUSION & FUTURE WORK

In conclusion, The model can be used with good faith in order to identify hate speeches online. The model is prone to type 1 errors, which does need further look at, yet it can identify real hate speech to great extents.
The future implementations regarding the model suggested in this paper, are explained as follows:

1. A third party software to govern hate speech and toxic comments on social media platforms in a user software.
   This model can be deployed in a software that takes input of the real-time comments on one's account and can alert the user regarding the same.
2. This model and procedure can be used with other datasets, focused on illicit possible messages on social media(as explored in the chapter 2.2)
   Illicit messages are more dangerous than some anonymous threats online, as these illicit messages are not threats, but actual attacks on society.
3. Corporate use. Used by social media platforms to attach to their automatic hate speech detection algorithm, this multi label classification might allow them to be able to generate a comprehensive report on whether a comment/post is toxic in nature or not.

REFERENCES

Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8).

Chen, J., Yan, S. and Wong, K., 2018. Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis. *Neural Computing and Applications*, 32(15), pp.10809-10818.

Raisi, E. and Huang, B., 2017. Cyberbullying Detection with Weakly Supervised Machine Learning.

Granizo, S. L., Caraguay, Á. L. V., López, L. I. B., & Hernández-Álvarez, M. (2020). Detection of possible illicit messages using natural language processing and computer vision on twitter and linked websites. *IEEE Access*, *8*, 44534-44546.

Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651*.

AI, C. (n.d.). *Toxic comment classification challenge*. Kaggle. Retrieved November 14, 2022, from https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data