Derogatory Comment Classification

Sudhanshu Chaurasia[1], KMR Dayaasaagar[2], Jayesh Girdhar[3] ,Dhanesh.S[4],Sunil Shelke[5]

[1,2,3,4] UG Student, Dept. of Computer Engineering, Pillai College of Engineering, New Panvel, India. [5] Assistant Professor, Dept. of Information Technology, Pillai College of Engineering, New Panvel, India.

---***---

Abstract - *Social media platforms have become a facet for people's opinion and reviews, wherein people share their opinions on a wide variety of topics. However, people exploit this facility to take a dig at those with whom they don't find their opinions match with. People use this facility to post harmful, racial, gender biased, threatful comments. As a result, social media platforms are quickly becoming indispensable. They often struggle to facilitate conversation, effectively forcing many communities to shut down user comments. This motivates us to look into this problem and build a model which will detect and classify derogatory comments. We will collect the dataset from Kaggle and experiment with the help of deep learning approaches like Naïve-Bayes, SVM, LSTM and BERT. algorithms. Thereby, we will classify the derogatory comments. We will compare these algorithms and conclude which algorithm is more effective.*

Key Words: **Language recognition, multilingual, offensive word, NBSVM**

1. INTRODUCTION

In today's current society, there is a big problem when it comes to online toxicity. Internet is an open discussing space for everyone to freely express their opinions. With the massive increase in social interactions on online social networks, there has also been an increase of hateful activities that exploit such infrastructure. This toxicity tends to negatively impact how a lot of people tend to engage in conversation and deters some from engaging in online conversation entirely.

As a result, online platforms tend to struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments. Harassment and abuse are discouraging people from sharing their ideas and disturbing the internet environment. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments if it's toxic. Motivated by this problem, we want to build a technology that detects and classifies the derogatory comment. For our model, the input will be a comment. We will use three different models to predict scores for "toxicity" (which is our target), "severe toxicity", "obscenity", "identity attack", "insult", "threat". Then Support Vector Machine (SVM) models are often used as baselines for other methods in text categorization .In this case Naive Bayes -SVM (NVB-SVM) provides more accuracy for further classification.

2. Literature Survey

A. Base-Line Model:Naive NVB-SVM was developed by Sida Wang and Christopher Manning. Bayes (NB) and Support Vector Machine (SVM) models are often used as baselines for other methods in text categorization .In this case Naive Bayes -SVM (NVB-SVM) provides more accuracy for further classification.

B. BERT: which stands for Bidirectional Encoder Representations transformers. Bidirectional Encoder Representations from Transformers (BERT) is a NLP model that was designed to pretrain deep bidirectional representations from unlabeled text and, after that, be fine-tuned using labeled text for different NLP tasks [7]. That way, with BERT model, we can create state-of-the-art models for many different NLP tasks [7]. We can see the results obtained by BERT in different NLP tasks at [7].

C. Toxic Comment Detection *using LSTM* This papers authors have developed a LSTM model to classify speech as hate or not with an accuracy of 95% accuracy. LSTM stands for long short-term memory and is an improvement over a normal RNN. The model not only classifies a given sentence as toxic or non-toxic but also gives the percentage of toxicity or non-toxicity of the given sentence.

D. Research on Text Classification Based on CNN and LSTM: In this paper a new model is developed by combining a LSTM and CNN which outperforms the performance of both the algorithms individually

2. 1 Summary of Related Work

The summary of methods used in literature is given in Table 1.

Literature	MODEL	Accuracy
Sida Wang, Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, 2020	NVB-SVM	87%
Hong Fu, Social Media Toxicity Classification Using Deep Learning ,2021	BERT	98%

Krishna Dubey, Toxic Comment Detection using LSTM ,2020	LSTM	95%
2019, Yuandong Luan Research on Text Classification Based on CNN and LSTM	LSTM-CNN	98%

3. Proposed Architecture

Our research is aimed at derogatory terms in a text that the user provides as input. Because social media has grown in popularity in recent years, there are people who engage in some type of online material (for example, articles) aimed towards a person or a group, etc. So, using natural language processing or machine learning methods, our program will evaluate the text and refer back to the dataset to determine the language of the input text as well as whether or not it contains any derogatory terms.

3.1 System Architecture

The proposed system architecture is given in Figure 1

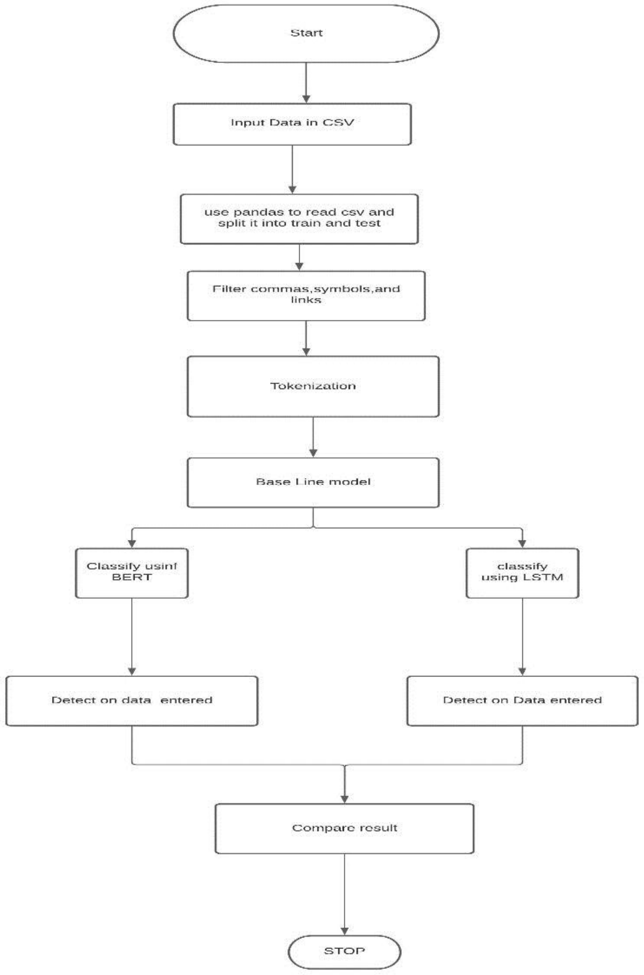


Fig. 1 Proposed system architecture

B. Block 1 Description: upload the data set which has been taken from Kaggle . It contains Wikipedia comments which is classified into Obscene, severely toxic, threat, identity hate and insult.

C. Block 2 Description: We will split the dataset into training and testing sets . we will split the dataset into 60% for training and 40% for testing and can be increased if there is a demand for increasing the accuracy.

D. Block 3 Description: This dataset contains data which has special, characters, useless words and links etc. We will be filtering them out .

E. Block 4 Description: Tokenization is an important step when working with text data. We will use this process to split the sentences into smaller units called tokens.

F. Block 5 Description: We will use Naïve Bayes-SVM model to create a base-line model so as to improve the accuracy in detecting and classifying the comments.

G. *Block 6 and 7 Description:* We will feed the output of the base line model to each of these models separately. BERT and LSTM will be able to classify and predict derogatory comments with higher accuracy rate.

H.Block 8 Description: We will compare the results generated by LSTM and BERT models in detecting derogatory comments .

3. Requirement Analysis

The implementation detail is given in this section.

3.1 Software

We will be using streamlit for making the web app. We will be using python for developing the classification model. We will be required to import different libraries to make the classification models .

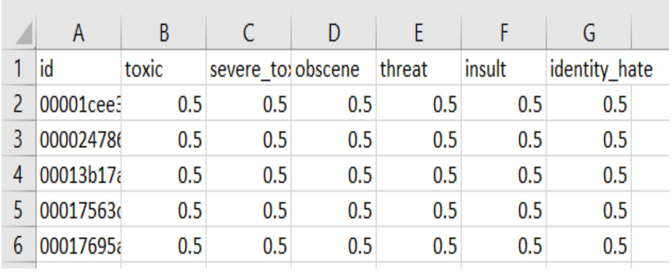
3.2 Hardware

We will be requiring Intel core i-5 generation as training and testing requires a high-performance processor. We will be requiring GeForce 1650 or above.

3.3 Dataset and Parameters

The Conversation AI team, a research initiative founded by Jigsaw and Google are working on tools to help improve online conversation. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the

Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content). A multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate is developed. An example of a data source by Jigsaw is given in Fig. 3.1.



	A	B	C	D	E	F	G
1	id	toxic	severe_to:	obscene	threat	insult	identity_hate
2	00001cee3	0.5	0.5	0.5	0.5	0.5	0.5
3	00002478(	0.5	0.5	0.5	0.5	0.5	0.5
4	00013b17a	0.5	0.5	0.5	0.5	0.5	0.5
5	00017563(	0.5	0.5	0.5	0.5	0.5	0.5
6	00017695a	0.5	0.5	0.5	0.5	0.5	0.5

Fig. 3.1 Kaggle Toxic dataset by jigsaw

3. CONCLUSIONS

To sum up our work, we implemented two machine learning models, namely, LSTM and BERT. We utilized information from data visualization to preprocess data. BERT can produce state-of-the-art work with only one output layer. We trained BERT with only one epoch and got results better than the LSTM model.

For future work, we would like to try ways to fix the problem of imbalanced data. BERT model brought us surprisingly good results and we would like to explore it if we have more time.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Dr Sunil Shelke for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to presenting this work.

REFERENCES

1.Sida Wang, Baselines and Bigrams: Simple, Good Sentiment and Topic Classification,2020

2. Hong Fu, Social Media Toxicity Classification Using DeepLearning,2021

3. Krishna Dubey, Toxic Comment Detection using LSTM, Third International Conference on Advances in Electronics , 2020

4. Yuandong Luan Research on Text Classification Based on CNN and LSTM,International Conference on Artificial Intelligence and Computer Applications,2019