Measurement of Toxicity Levels in Sentences using NLP and CNN with Word Embedding System

Mahmudul Hasan Shakil, Sazal Kanti Kundu,
Zawad Alam, Arnob Kumar Dey
Department of Computer Science and Engineering, Brac University
Dhaka,Bangladesh.
{mahmudul.hasan.shakil, sazal.kanti.kundu, zawad.alam, arnob.kumar.dey }@g.bracu.ac.bd

Abstract

We are living in an era of technology. But unfortunately, it has also become a place for hateful activity, abusive words, cyber-bullying, and anonymous threats. In this paper, we are using two basic neural networks which are Convolutional Neural Networks (CNN) and Natural language processing (NLP) as a concern for reducing toxicity in social platforms.

1 Introduction

The rapid growth of information technology and the disruptive transformation of social media have happened in recent years. Websites like Facebook, Twitter, Instagram, where people can express their thoughts or feelings by posting text, photos or videos, have become incredibly popular. But it has also a bad reputatin for spreading toxicity. There are many existing works in this field but those are not fully successful yet to provide accuracy in satisfactory level. In this paper, we employ natural language processing (NLP) and convolution neural network (CNN).

2 Related Works

Georgakopoulos, S. V. et al. used CNN model for solving toxic comment classification problem using the same dataset we have used in our methodology.Saeed, H. H. et al applied deep neural network architectures with a good accuracy.Furthermore, Kandasamy et al adopted natural language processing technique (NLP) integrated with the implementation of URL analysis and supervised machine learning techniques social media data where it scored 94 % accuracy. Anand, M. et al presented different deep learning techniques such as Convolution neural network (CNN), ANN, long short term memory cell (LSTM) and these are with and without word GloVe embeddings, where GloVe pre-trained model is applied for classification.

3 Methodology

The dataset we have used in our research, is acquired from Kaggle which is very popular publicly available dataset named "Wikipedia Talk Page Comments annotated with toxicity reasons".

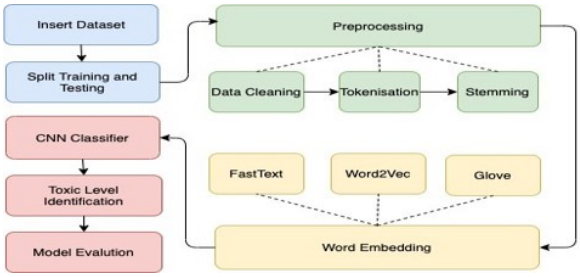

Figure 1: Proposed Model

We have designed our model which is showing in Figure 1. First of all, we will split the dataset into test and training set. Then, we will apply data cleaning to remove such words like I, is, are etc. After that, we will use tokenization to vectorize the word into numerical number with some values. We will also use stemming to reduce repetitive words like their meaning is same but their form of tense is different. For text classification specifically the utilization of Convolutional Neural Networks (CNN) have as of late been proposed moving toward text examination in an advanced way underlining in the structure of words in a record.

4 Conclusion

In this paper, we have presented a toxic comment classification system which is an essential tool for social media sites. With the ever-expanding popularity and use of social media platforms, the numbers of vulgar and negative comments are also increasing. The system is also imperative to preclude the cyber bullying, toxic or offensive comments as addressing these issues are still grueling.