HATE SPEECH DETECTION AND SENTIMENT ANALYSIS

Mansi Dhawan, Dr. M.L. Sharma

[1]*Student, Maharaja Agrasen Institute of Technology*
[2]*H.O.D (I.T.), Maharaja Agrasen Institute of Technology*

Abstract - *One of the major breakthroughs in internet is of social media and micro blogging websites. It acted as a platform for people to express their views, opinions on a topic or various aspects in life. This motivated me to perform sentiment analysis and hate speech detection on such a dynamic corpus amount of data available out there.*

Sentiment Analysis has become essential business wise as well socially so as to analyze how millions of people take in the information and changes happening around the world and how it affects their lives. With growing popularity of social media and the anonymity and convenience it offers, has led to increase in hate speech, therefore, there is an urgent need for effective solution or countermeasures to tackle this problem

In my paper, I have performed sentiment/emotion analysis on audio and recognize various emotions such as happy, sad, calm, angry etc. Also, I have performed hate speech detection on tweets, YouTube videos and comments for detection of the same by using various deep learning methods and algorithms.

Key Words: **Sentiment analysis, Hate Speech Detection, LSTM, CNN, NLP, BERT, Speech Emotion Recognition, Deep Learning**

1. INTRODUCTION

1.1 Sentiment Analysis and SER

Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language. Traditionally, sentiment analysis has been about opinion polarity, i.e., whether someone has positive, neutral, or negative opinion towards something.

The object of sentiment analysis has typically been a product or a service whose review has been made public on the Internet. Although, I think it is more accurate to view sentiments as emotionally loaded opinions. .

Speech Emotion recognition (SER) is gaining enormous popularity. This project attempts to use deep learning method Convolutional Neural Network (CNN) to recognize emotion and classify the emotion according to the speech signals.

1.2 Hate Speech Detection

The term 'hate speech' is formally defined by oxfords as 'abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation'. Hatred is generally based on ethnicity, religion, disability, gender, caste, and sexual orientation

The internet and social media has become a powerful tool for such propagandist to spread hate and reach new audience. The anonymity and flexibility that the internet offers allow such haters to easily and safely propagate hate without any fear. Lack of regulation and legal policy worsens the situation a bit more. This is largely because such measures involve manual, labour intensive and inefficient method of identification and removal of offensive materials.

The need of the hour is for automated state of the art and scalable methods for detection and classification methods.

In this research paper, I have attempted to perform binary as well has multi label hate speech detection on platforms, Twitter (tweets) and YouTube (comments and videos), through deep learning algorithms like CNN, BI-LSTM etc

2. RESEARCH METHODOLOGY AND FRAMEWORK

2.1 Research Methodology

1. Collect data using twitter API and store them in a data frame i.e. in a tabular form. Then store the data frames in csv format.
2. For YouTube comments and transcripts, develop a simple scraper that crawls a YouTube video's comment page and the transcripts. The crawler uses Ajax to go through every comment and transcripts on a YouTube video and then saves them to separate json file. Store the json file in csv format.
3. Preprocess the data by removing special character, user name, links emojis etc and perform tokenization, stemming.
4. Using Convolutional Neural Network to recognize emotion/sentiments from the audio recording to identify emotions like sad, happy, calm, surprised, disgusted, angry etc
5. Hate speech detection on multimedia is performed as follows:

5.1 Binary label classification of textual data using BI-LSTM
5.2 Multi label classification of textual data using BERT CNN into different categories like threat, insult, obscene, identity hate etc
5.3 Binary label classification in YouTube videos using BI LSTM.

2.2 Research Framework

Twitter API: Twitter is an information network and communication mechanism that produces more than millions of tweets a day. The Twitter platform offers access to that corpus of data, via the APIs. Each API represents a facet of Twitter, and allows developers to build upon and extend their applications in new and creative ways. The Twitter API allows you to access the features of Twitter without having to go through the website interface. This can be useful for doing things like posting tweets or sending directed messages in an automated way with scripts.

Long Short-Term Memory (LSTM): These networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.

An LSTM layer consists of a set of recurrently connected blocks, known as memory blocks. These blocks can be thought of as a differentiable version of the memory chips in a digital computer. Each one contains one or more recurrently connected memory cells and three multiplicative units – the input, output and forget gates – that provide continuous analogues of write, read and reset operations for the cells. [1] BI LSTM means bidirectional LSTM, which means the signal propagates backward as well as forward in time.

Convolutional Neural Network (CNN): It is one of the variants of neural networks used heavily in the field of Computer Vision. It derives its name from the type of hidden layers it consists of. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers, and normalization layers [2]

Bidirectional Encoder Representations for Transformers (BERT): It is a deep learning model that has given state-of-the-art results on a wide variety of natural language processing tasks. It stands for Bidirectional Encoder Representations for Transformers. It has been pre-trained on Wikipedia and Books Corpus and requires task-specific fine-tuning.

BERT is a multi-layer bidirectional Transformer encoder. There are two models:

BERT base – 12 layers (transformer blocks), 12 attention heads, and 110 million parameters.

BERT Large – 24 layers, 16 attention heads and, 340 million parameters. [3]

TensorFlow: TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

TensorFlow allows developers to create *dataflow graphs*—structures that describe how data moves through a graph, or a series of processing nodes. Each node in the graph represents a mathematical operation, and each connection or edge between nodes is a multidimensional data array, or *tensor*.[4]

3. CONCLUSIONS

From the result obtained and charts visualized in performing binary label classification on twitter and YouTube comment, following is concluded:

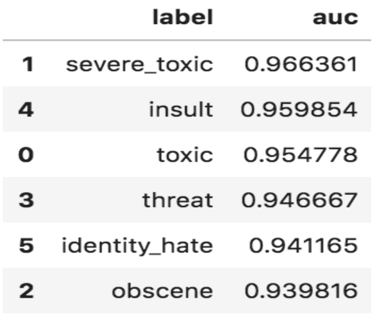
Input: Training dataset of 32,000 youtube comments and tweets.

- Accuracy : 89%
- Accuracy in hate comments :33%
- Accuracy in non hate comments: 96%

From the result obtained and charts visualized in performing multi label classification on twitter and YouTube comment, following is concluded:

Input –160000 comments and tweets training data set for multiclass classification.

Comments are scored in six categories – toxic, obscene, threat, insult, identity hate, severe toxic.



	label	auc
1	severe_toxic	0.966361
4	insult	0.959854
0	toxic	0.954778
3	threat	0.946667
5	identity_hate	0.941165
2	obscene	0.939816

Fig -1: Accuracy for different categories

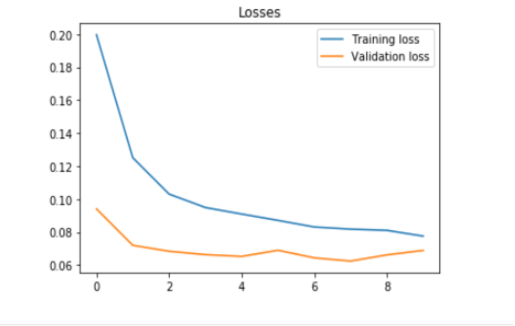


Chart -1: Training and validation loss plot

Key Take Away From Sentiment/ emotion recognition on audios:

Input: 24 different actors recorded speech and song version (Source -The Ryerson Audio-Visual Database of Emotional Speech and Song).

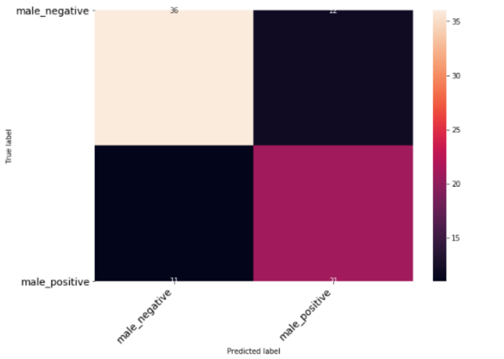


Fig -2: Confusion Matrix

- Accuracy achieved : 66.88
- F1–score : 0.66
- Emotions are subjective and it is hard to notate them.
- We should define the emotions that suitable for own project objective.
- Deciding the input for your model as a sentence, a recording or an utterance is a difficult task.
- It is complex and very expensive to build a good speech emotion dataset.

From the result obtained and charts visualized in performing hate speech classification of YouTube video transcripts, following is concluded:

Input: Video transcripts/captions scraped for different videos and stored in csv format.

Training set: Same as used for hate speech detection for comments and tweets.

- Accuracy : 86.9%
- Accuracy in hate comments :42%
- Accuracy in non hate comments: 92%
- F1 - score: 0.41

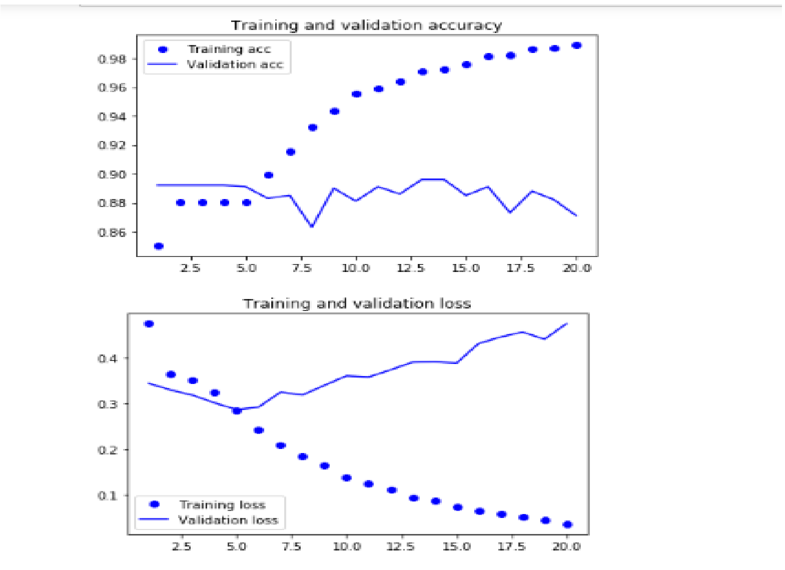


Chart -2: Training and validation accuracy and loss plot

Future Work

As hate speech continues to be a societal problem and sentiment analysis is becoming a business as well as a social need, the need for automatic hate speech detection and sentiment analysis systems become more apparent. I presented the current approaches for this task as well as a new system that achieves reasonable accuracy.

However there are a few challenges and future scope that were faced that need to be addressed:

- Only English language was analysed while the corpus data available is not restricted to English language.
- Predicting using more features than just the words present in the text.
- To produce human like efficiency and accuracy in detecting the hate speech or the sentiment.
- Spelling mistakes and abbreviations cause hinder in producing accurate results.
- Analysis of videos without transcripts.
- Images as a data set was not used in sentiment or hate speech analysis
- Really short videos and audio recordings could be used as high storage and processing capacity is used for efficient results in case of longer recordings.
- Preprocessing the data like cropping silence voice, normalizing the length by zero padding, etc.
- Experimenting with other algorithms and approaches on this topic.

REFERENCES

[1] https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

[2] https://towardsdatascience.com/understanding-neural-networks-from-neuron-to-rnn-cnn-and-deep-learning-cd88e90e0a90

[3] https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/

[4] https://www.tensorflow.org/

[5] https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd

[6] H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan.

[7] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. 2017

[9] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. 2017.

[10] Automated Hate Speech Detection and the Problem of Offensive LanguageThoma Davidson Dana Warmsley, Michael Macy, Ingmar Weber

[11] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recogn