HATE SPEECH DETECTION AND PREVENTION USING DEEP LEARNING

PHASE I REPORT

Submitted By

JOSHUA S – 2019202019

in partial fulfilment for the award of the degree of

MASTER OF COMPUTER APPLICATION

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY, CHENNAI - 600025

APRIL, 2022

ANNA UNIVERSITY, CHENNAI

BONAFIDE CERTIFICATE

Certified that this report titled "**HATE SPEECH DETECTION USING DEEP LEARNING**" is the bonafide work of **JOSHUA S (2019202019)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

DR. S. SRIDHAR
Professor and Head of the Department
Department of IST
Anna University
Chennai –600025

MS. B. SIVA SHANKARI
Project Guide
Department of IST
Anna University
Chennai – 600025

ABSTRACT

Hate speech is a subject of worry for online platforms. Hate speech is a form of expression that assaults a person or a community based on race, origin, religion, sexual orientation, or other attributes. Although it can be expressed in multiple ways, both online and offline, the increasing popularity of social media has exponentially increased both its use and severity. With powerfully expanding datasets manual mediation of posts is very inconceivable or will be tedious. Hate speech detection should be an automated task to distinguish hate speech from the provided input. This work combines data analysis and natural language processing strategies, to sensitize all social media providers to the pervasiveness of hate on social media. Specifically, we use sentiment and emotion analysis algorithms to find the emotion & sentiment in the text and a deep learning model multi-channel convolutional neural network (MCCNN) is implemented to detect hate speech. The model consists of 3 channels of Convolutional Neural Network (CNN). Each channel is merged and connected to a fully connected layer from where the final output is obtained. The model is trained by using the Hate Speech and Offensive Language Dataset. With the use of the model, the inputs of a comment section are verified for hate text and prevents the user from posting the comment online, until or unless he/she changes the text which doesn't have any hate speech.

ACKNOWLEDGEMENT

The satisfaction that accompanies the success would be incomplete without mentioning the names of people who made it possible.

I would like to express my earnest thanks to **Ms. B. Siva Shankari**, Teaching Fellow, Department of Information Science and Technology, CEG, Anna University for her valuable guidance, encouragement and attitude that has driven the project work in a steady pace to a successful completion.

I would like to thank **Dr. S. Sridhar**, Professor and Head, Department of Information Science and Technology, CEG, Anna University for his kind support.

I express my sincere thanks to the project committee, **Dr. Saswati Mukherjee,** Professor, Department of Information Science and Technology, **Dr. M. Vijayalakshmi**, Associate Professor, Department of Information Science and Technology, **Dr. E. Uma,** Assistant Professor, Department of Information Science and Technology, **Ms. P. S. Apirajitha**, Teaching Fellow, Department of Information Science and Technology, **Ms. C. M. Sowmya**, Teaching Fellow, Department of Information Science and Technology, for their valuable suggestions that have led to the betterment of the project.

Joshua S

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	iii
	ACKNOWLEDGEMENT	iv
	LIST OF FIGURES	vi
1	**INTRODUCTION**	**1**
	1.1 GENERAL	1
	1.2 NEED FOR THE STUDY	1
	1.3 PROBLEM STATEMENT	2
	1.4 MOTIVATION & OBJECTIVES	2
	1.5 DOMAIN	2
2	**LITERATURE REVIEW**	**4**
3	**SYSTEM ARCHITECTURE**	**6**
	3.1 ARCHITECTURE DIAGRAM	6
	3.2 ARCHITECTURE EXPLANATION	7
	3.3 LIST OF MODULES	7
	3.4 MODULES EXPLANATION	7
4	**IMPLEMENTATION**	**10**
	4.1 PLATFORM / FRAMEWORK	10
	4.2 MODULES ALGORITHM	10
	4.3 MODULE BASED SCREENSHOTS	13
5	**REFERENCES**	**18**

LIST OF FIGURES

S.NO	TITLE	PAGE NO.
3.1	Hate Speech Detection	6
3.2	Data Acquisition & Pre-processing	8
3.3	Sentiment Analysis	8
3.4	Emotional Analysis	9
3.5	MCCNN Model	9

CHAPTER – 1
INTRODUCTION

1.1 GENERAL

Social mediums are a platform to express one’s voice and should be a place to connect freely without any fear of attack. One can define hate speech as expressions of hatred or encouraging violence towards a group of individuals or individuals based on race, religion, sex, sexual orientation, illness, disability, national origin, gender, colour, etc. This definition varies from one country to the other. It includes various forms of expression that spread false information or promotes or justifies violence, hatred, discrimination against a person or group of persons for several reasons. Usage of offensive and abusive words for people based on any characteristics can harm one’s peace and might lead to depression.

1.2 NEED FOR THE STUDY

It is necessary to eradicate this problem and demands a tool to rectify it in online platforms. If it remains unaddressed it can lead to serious crimes, violence, and conflicts to a great extent. It’s a type of activity that cannot be tolerated and contributes to crime. It sets up a link between violence and hates speech. It has had a significant impact on the Internet community, as a number of people, including well-known figures, have been targeted by explicit 'harassers.' These harsh comments are likely to have an impact on the targets, leading to demoralization. Therefore, there must be some tools to detect hate speech on different platforms that are spoiling the social environment. Detection of hate speech and its removal can not only maintain a positive environment in social media but can also prevent the crime rate.

1.3 PROBLEM STATEMENT

Detecting the Hate Speech that has been spread online using a 3 channel Multi-Channel Convolutional Neural Network (MCCNN) which is a Deep Learning method and prevent it using suitable methods, such that a user will be restricted to post the comment online.

1.4 MOTIVATION & OBJECTIVES OF THE STUDY

Motivation

Hate speech is a subject of worry for online platforms. Cursing, considering women as objects, comments on physical appearance, inferiority, comparisons, generalization, mocking any events, etc. Hate speech detection should be an automated task to distinguish hate speech from the provided input.

Objectives

To understand the research carried out by various other researchers in the same field and their implementation methodologies. And understanding the pre-requisite study and to get hands on experience in Machine Learning, Deep Learning and Natural Language Processing (NLP) Classifiers.

Then, to build an efficient system that identifies hate speech and abusive language in the comments, such that it serves a noble purpose for mankind.

1.5 DOMAIN

Deep Learning

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. In other words, Deep Learning is a subset of Machine Learning, which on the other hand is a subset of Artificial Intelligence. Artificial Intelligence is a general term that refers to techniques that enable computers to mimic human behaviour. Machine Learning represents a set of algorithms trained on data that make all of this possible.

Deep learning is a key technology behind driverless cars, enabling them to recognize a stop sign, or to distinguish a pedestrian from a lamppost. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It’s achieving results that were not possible before.

In deep learning, a computer model learns to perform classification tasks directly from images, text, or sound. Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers.

CHAPTER – 2

LITERATURE REVIEW

T. A. Naidu and S. Kumar [1]. Proposed that Hate speech detection should be an automated task to distinguish hate speech from the provided input. In the work, they have implemented a deep learning model multi-channel convolutional neural network (MCCNN). The model consists of 3 channels of Convolutional Neural Network. Each channel is merged and connected to a fully connected layer from where the final output is obtained. The model is compared with a single-channel convolution neural network and the results have shown that MCCNN outperformed simple CNN. The accuracy and F1-score achieved by the model are 95.49 and 93.93.

Many scientists had worked on machine learning models that combined traditional and deep learning techniques. Studies have shown that deep learning models have shown tremendous results for the detection of hate speech. Starting with some traditional machine learning models, SVM classifiers for hate speech detection and got some good results, naive base classifier for the classification of tweets, logistic regression (LR) for detection and achieved a precession of 90%. Some focused on two words racism and sexism, to detect hate speech on Twitter; a multi-class classifier to recognize offensive and hate language in tweets. Some methods used Recurrence Neural Network with uni-gram and bi-gram character embedding for hate speech detection. Others used word2vec word vectors and CNN as a classifier which gave some good results.

A. Rodriguez, Y. -L. Chen and C. Argueta [2]. Proposed that the aim of their research is to locate and analyze the unstructured data of selected social media posts that intend to spread hate in the comment sections. To address the issue, they have proposed a novel framework called FADOHS, which combines data analysis and natural language processing strategies, to sensitize all social media providers to the pervasiveness of hate on social media. Specifically, they have used sentiment and emotion analysis algorithms to analyze recent posts and comments on these pages.

Posts suspected of containing dehumanizing words will be processed before fed to the clustering algorithm for further evaluation. According to the experimental results, the proposed FADOHS framework is able to surpass the state-of-the-art approach in terms of precision, recall, and F1 scores by approximately 10%.

The resulting framework identified a set of sensitive topics that can promote hate. The major contributions of this study are as follows. First, they developed a semiautomatic method to discover pages that discuss sensitive topics. Second, we propose an automatic method to cluster posts from pages that discuss specific topics. Finally, they designed and implemented a new framework for hate speech detection.

K. A. Qureshi and M. Sabih [3]. Addressing different categories of hate separately, this paper aims to accurately predict their different forms, by exploring a group of text mining features. Two distinct groups of features are explored for problem suitability. These are baseline features and self-discovered/new features. Baseline features include the most commonly used effective features of related studies. Exploration found a few of them, like character and word n-grams, dependency tuples, sentiment scores, and count of 1st, 2nd person pronouns are more efficient than others. Due to the application of latent semantic analysis (LSA) for dimensionality reduction, this problem is benefited from the utilization of many complex and non-linear models and CAT Boost performed best.

CHAPTER – 3

SYSTEM ARCHITECTURE

3.1 ARCHITECTURE DIAGRAM

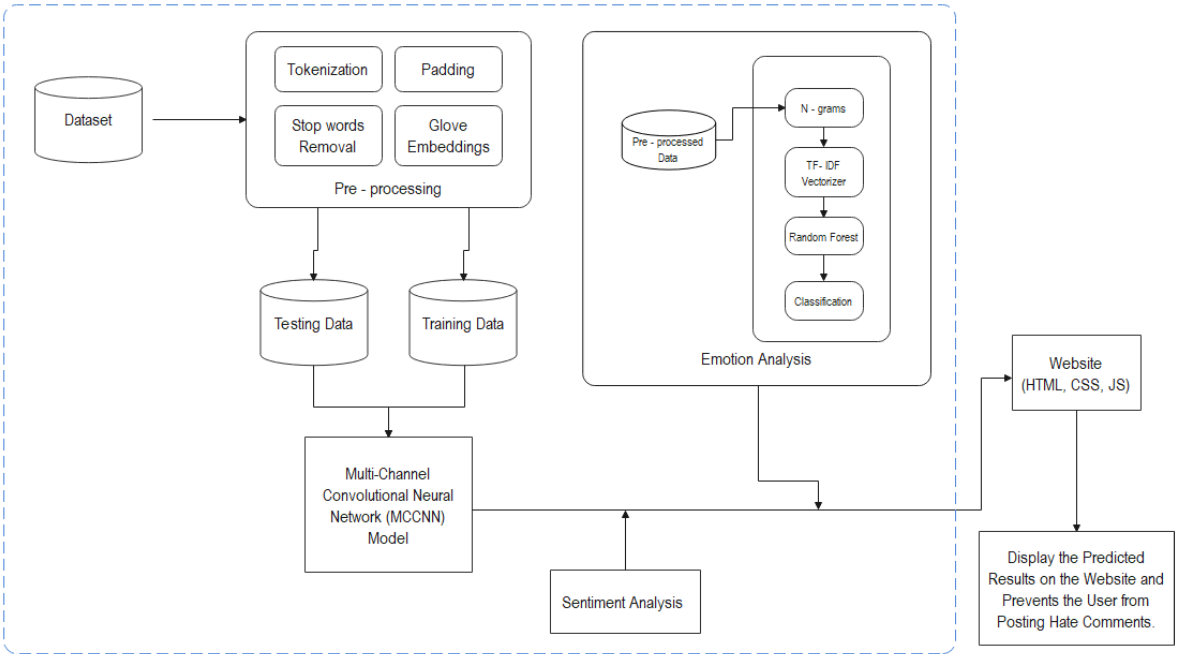


Fig 3.1 *Hate Speech Detection*

3.2 ARCHITECTURE EXPLANATION

The dataset is obtained from Kaggle which contains hate tweets. The data are pre-processed by removing noise from the dataset. Glove embeddings is performed on the tokenized data to create a co-occurrence matrix. The Data is split into training and testing for the model. Multi-Channel Convolutional Neural Network (MCCNN) is used to create the model to predict Hate speech. The MCCNN has 3 channels to improve the accuracy of the model to predict the Hate Speech in text. Then Sentiment and Emotional analysis are performed to understand the meaning of tweets. Once the model is done, it will be connected to a website, which takes an input and classifies it as Hate speech, if there is hate speech in a statement or comment and prevents it from being posted online.

3.3 LIST OF MODULES

1. Data acquisition & Pre-processing
2. Sentiment
3. Emotion analysis
4. Multi-Channel Convolutional Neural Network (MCCNN) model
5. Front end

3.4 MODULES EXPLANATION

Data acquisition & Pre-processing:

The dataset for predicting the hate speech is obtained from Kaggle to train and test the model. The dataset consists of hate speech which is derived from tweets from twitter. The obtained data is processed before using it to train the model. Tokenization, Stemming, Lemmatization, Stop words removal and Chunking (Natural Language Processing) are done on the dataset to reduce the noise overfitting of the model.

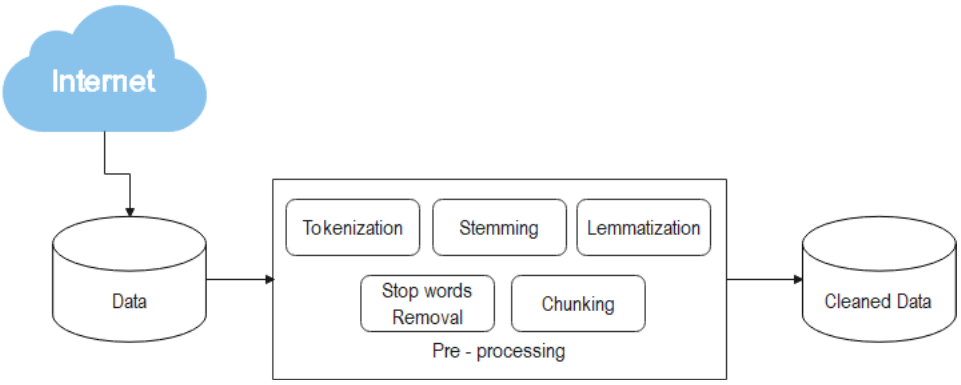


Fig 3.2 *Data Acquisition & Pre-processing*

Sentiment Analysis:

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral is defined as Sentiment analysis.

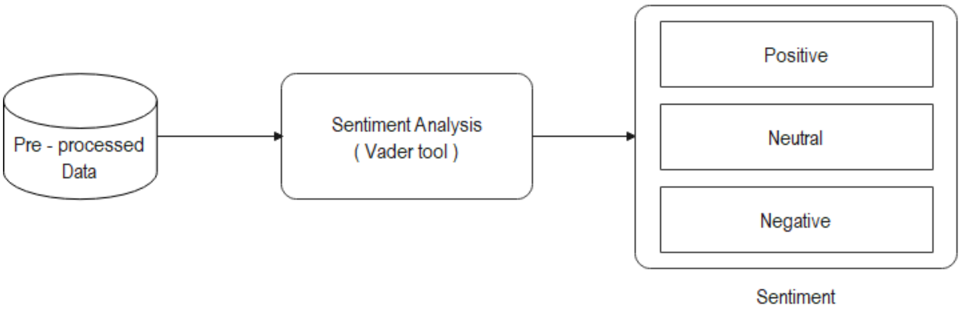


Fig 3.3 *Sentiment Analysis*

Emotion analysis:

Emotion analysis is the process of identifying and analysing the underlying emotions expressed in textual data. Where the Pre – processed data is split using uni-grams and bi-grams and vectorized by Term Frequency – Inverse Document Frequency (TF-IDF) and Random Forest model is trained and tested to classify the emotions.

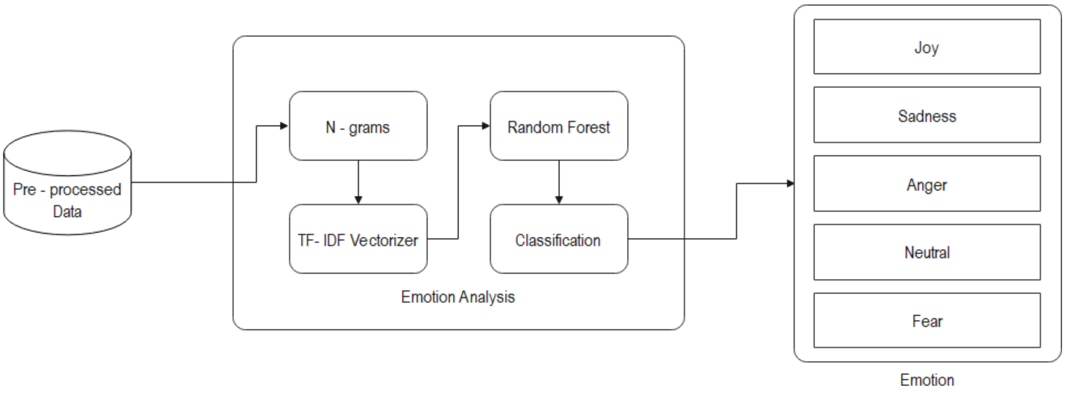


Fig 3.4 *Emotional Analysis*

Multi-Channel Convolutional Neural Network (MCCNN):

MCCNN is a CNN with more than one channel which helps the neural network to achieve more accuracy and precision. This model is used to predict the Hate Speech in the comment.

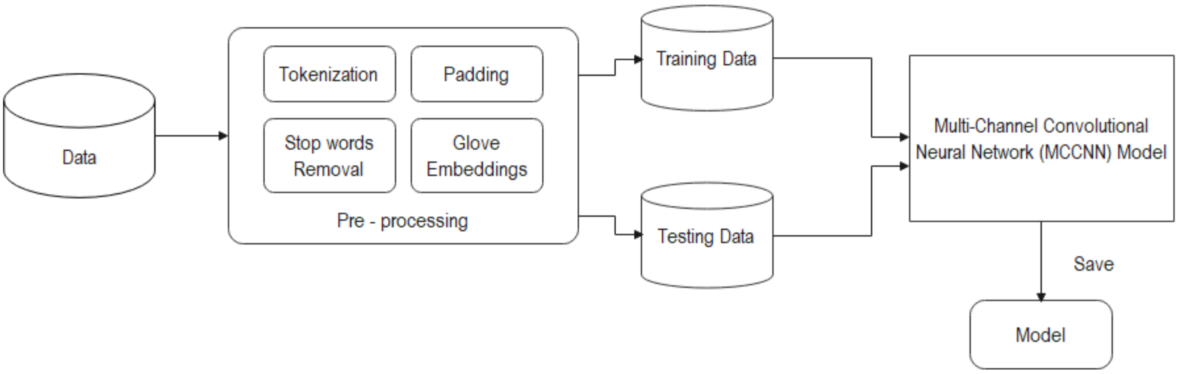


Fig 3.5 *MCCNN Model*

Front end:

Finally, a website is used to visualize the hate speech detection and prevention, which is developed by HTML, CSS and JS.

CHAPTER – 4

IMPLEMENTATION

4.1 PLATFORM / FRAMEWORK

Tools:

1. Windows OS
2. Python
3. VS code
4. Jupyter Extension / Jupyter Notebook
5. Tenserflow
6. Keras
7. Nltk
8. Sklearn

4.2 MODULES ALGORITHM

Data Acquisition and Pre-processing

```
Step1: Start
Step2: Import numpy and pandas
Step3: remove stop words, symbols, punctuations, etc
Step4: Convert all text to lower case
Step5: Stop
```

Sentiment Analysis

Step1: Start

Step2: Import nltk

Step3: Import Vader from nltk.sentiment

Step4: Use SentimentIntensityAnalyzer() from Vader

Step5: Check the polarity score of the text/comment

Step6: Stop

Emotion Analysis

Step1: Start

Step2: Download dataset

Step3: Import the required modules

Step4: pre-process to remove noise from the data

Step5: Convert the text data into vector by performing TF-IDF vectorizer

Step6: Create a Random Forest Classifier

Step7: Train and Test the model

Step8: Check the Accuracy, Precision and Confusion Matrix

Step9: Save the model to use it for prediction

Step10: Stop

CNN

Step1: Start

Step2: Import tensorflow and keras

Step3: Add Glove embedding to the tokenized data

Step4: Split the data into train and test

Step5: Create a convolutional base layer with Conv1D with relu activation

Step6: Add Max pooling layer on top

Step7: Add Dense layer with relu activation

Step8: Add Flatten layer

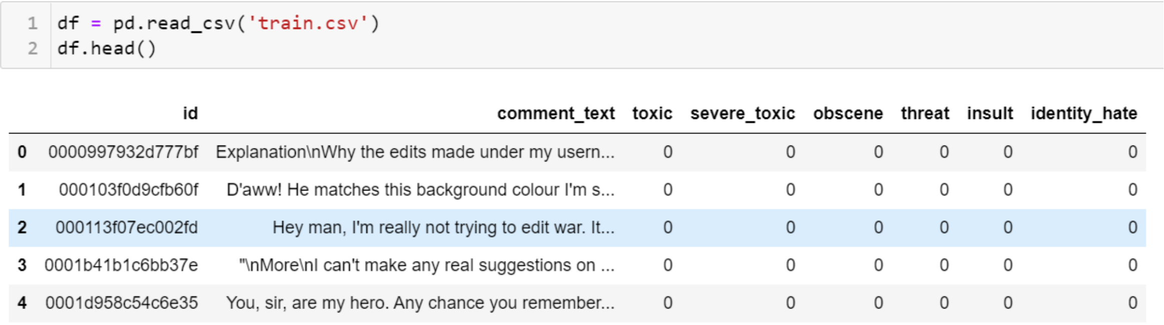
Step9: Add Dense layer with sigmoid with sigmoid activation

Step10: Compile the model with binary cross entropy and adam optimizer

Step11: Stop

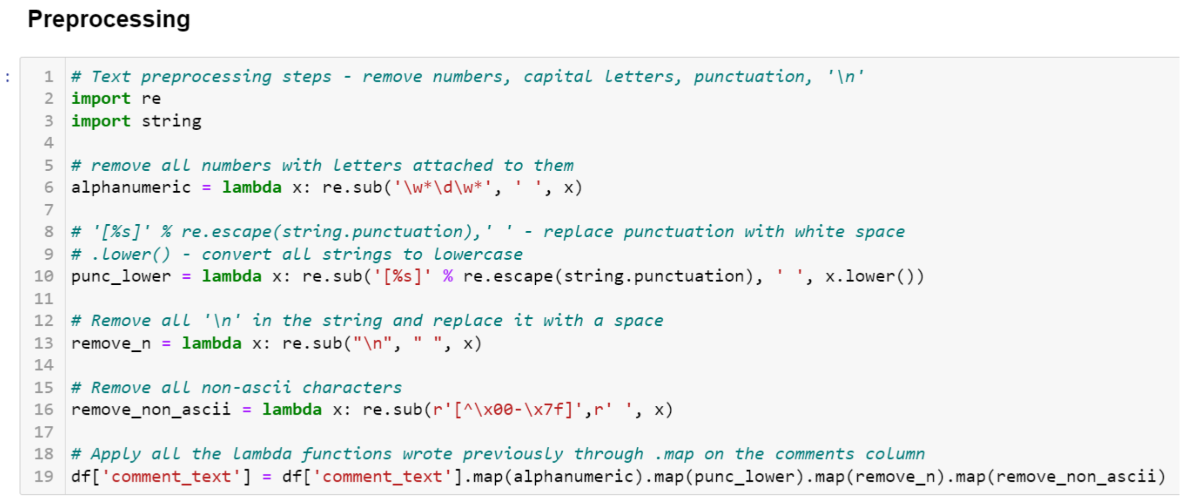
4.3 MODULE BASED SCREENSHOTS

Data acquisition & Pre-processing



```
df = pd.read_csv('train.csv')
df.head()
```

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

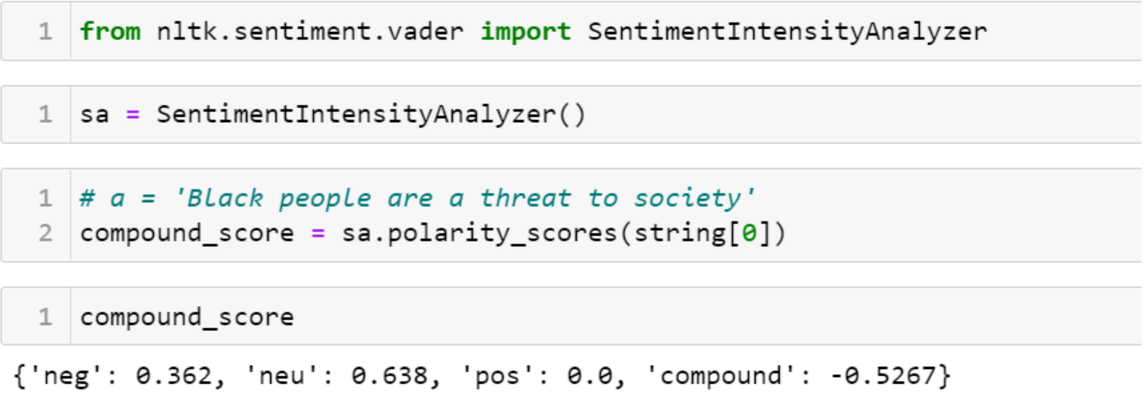


Preprocessing

```
# Text preprocessing steps - remove numbers, capital letters, punctuation, '\n'
import re
import string

# remove all numbers with letters attached to them
alphanumeric = lambda x: re.sub('\w*\d\w*', ' ', x)

# '[%s]' % re.escape(string.punctuation),' ' - replace punctuation with white space
# .lower() - convert all strings to lowercase
punc_lower = lambda x: re.sub('[%s]' % re.escape(string.punctuation), ' ', x.lower())

# Remove all '\n' in the string and replace it with a space
remove_n = lambda x: re.sub("\n", " ", x)

# Remove all non-ascii characters
remove_non_ascii = lambda x: re.sub(r'[^\x00-\x7f]',r' ', x)

# Apply all the lambda functions wrote previously through .map on the comments column
df['comment_text'] = df['comment_text'].map(alphanumeric).map(punc_lower).map(remove_n).map(remove_non_ascii)
```

Data is loaded and pre-processed to remove numbers, stop word, etc. Which helps in reducing overfitting of the model resulting in an increase in accuracy of the model.

Sentiment Analysis



```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
```

```
sa = SentimentIntensityAnalyzer()
```

```
# a = 'Black people are a threat to society'
compound_score = sa.polarity_scores(string[0])
```

```
compound_score
```

```
{'neg': 0.362, 'neu': 0.638, 'pos': 0.0, 'compound': -0.5267}
```

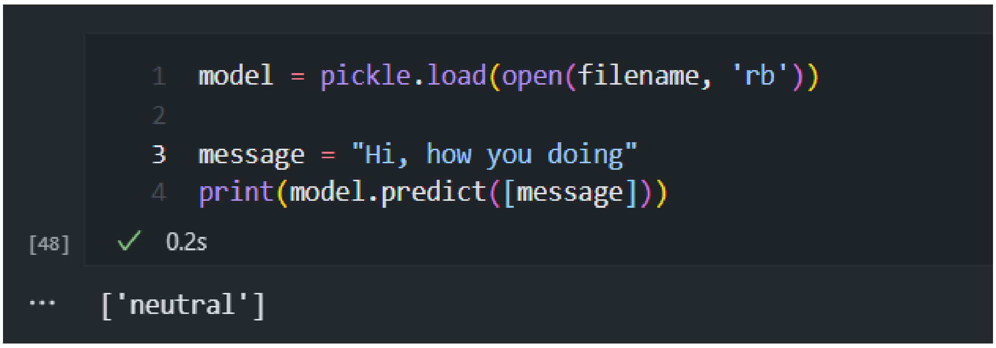
The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive).

$$x = \frac{x}{\sqrt{x^2 + \alpha}}$$

where x = sum of valence scores of constituent words, and

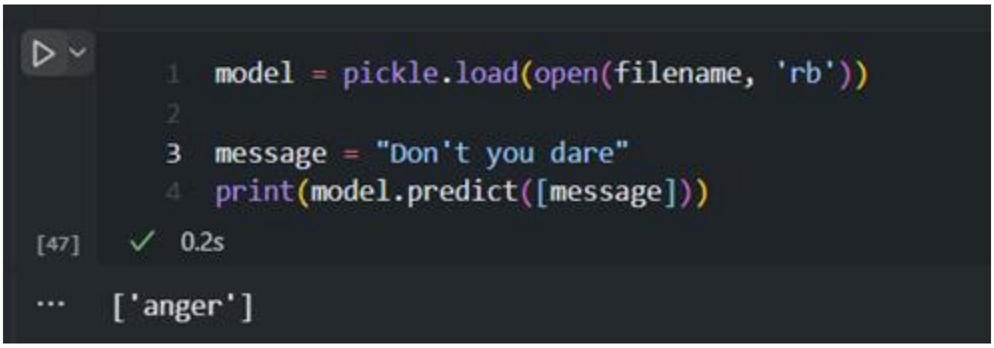
α = Normalization constant (default value is 15)

Emotional Analysis



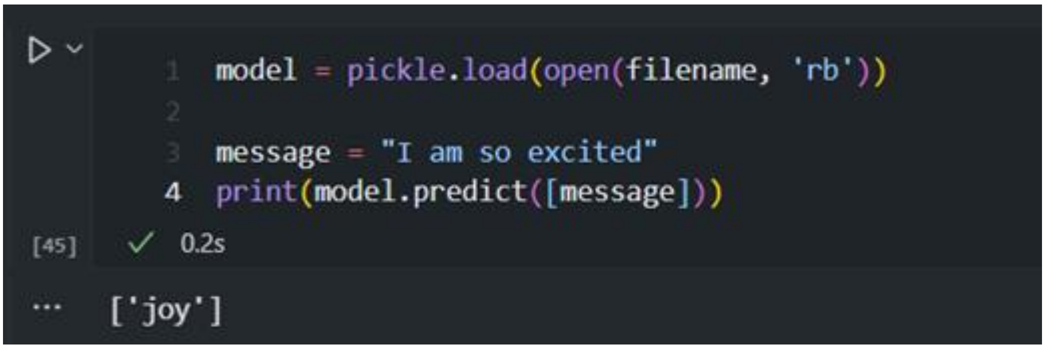
```
model = pickle.load(open(filename, 'rb'))

message = "Hi, how you doing"
print(model.predict([message]))
```

```
['neutral']
```



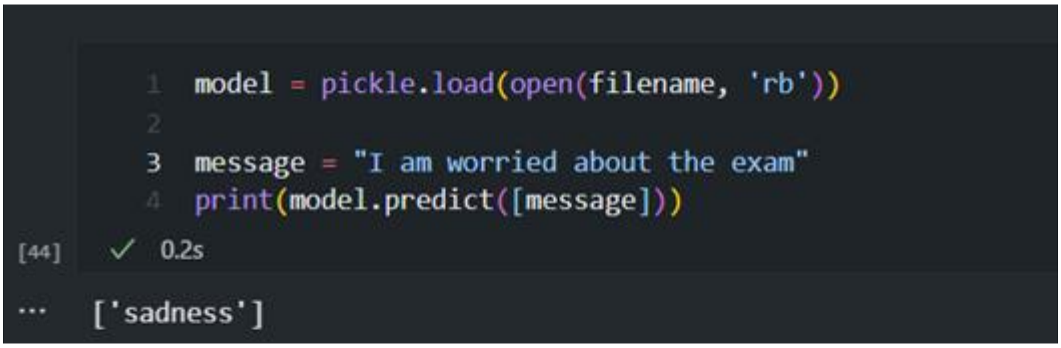
```
model = pickle.load(open(filename, 'rb'))

message = "Don't you dare"
print(model.predict([message]))
```

```
['anger']
```



```
model = pickle.load(open(filename, 'rb'))

message = "I am so excited"
print(model.predict([message]))
```

```
['joy']
```

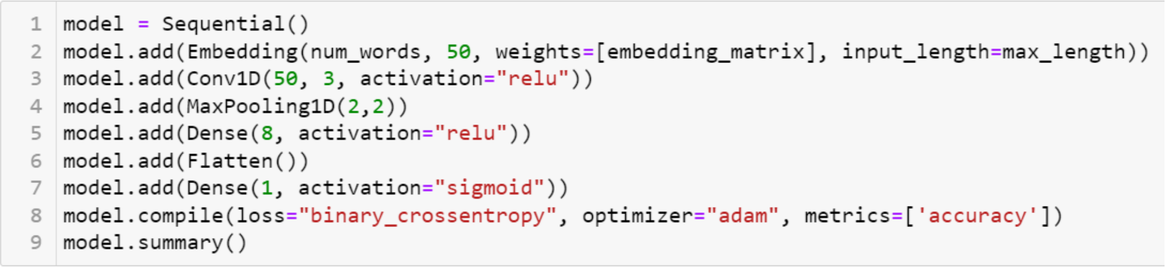


```
model = pickle.load(open(filename, 'rb'))

message = "I am worried about the exam"
print(model.predict([message]))
```

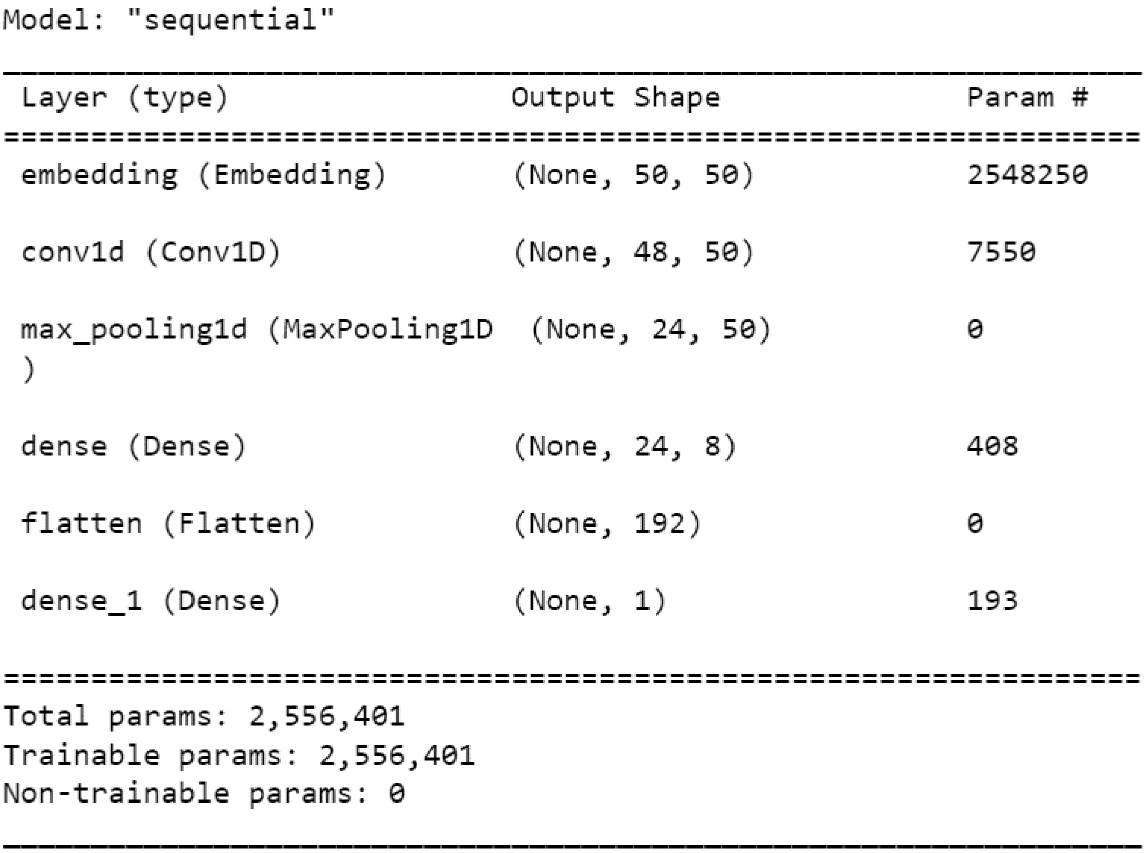
```
['sadness']
```

The Emotional Analysis model predicts the emotion based on the given input from one of the five emotions Joy, Sadness, Fear, Anger, Neutral.

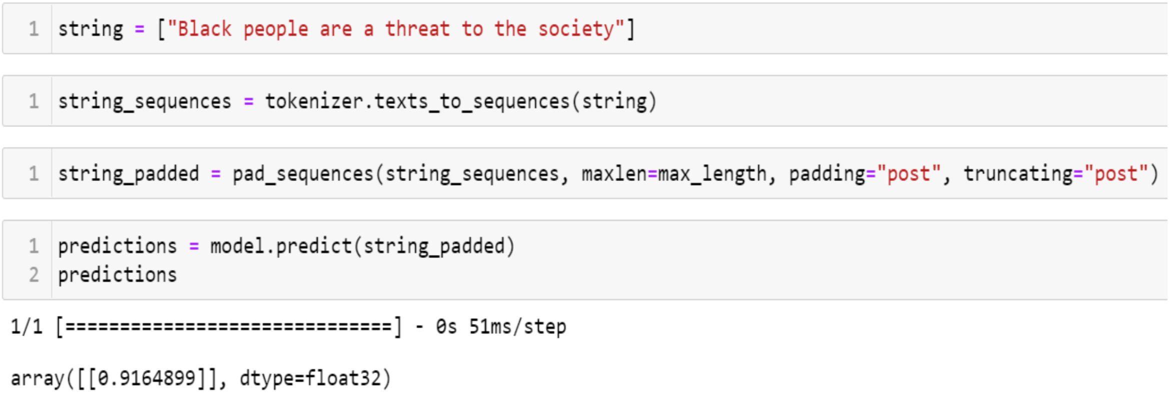
Hate Speech Detection using CNN



```python
model = Sequential()
model.add(Embedding(num_words, 50, weights=[embedding_matrix], input_length=max_length))
model.add(Conv1D(50, 3, activation="relu"))
model.add(MaxPooling1D(2,2))
model.add(Dense(8, activation="relu"))
model.add(Flatten())
model.add(Dense(1, activation="sigmoid"))
model.compile(loss="binary_crossentropy", optimizer="adam", metrics=['accuracy'])
model.summary()
```



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 50, 50)            2548250

 conv1d (Conv1D)             (None, 48, 50)            7550

 max_pooling1d (MaxPooling1D  (None, 24, 50)           0
 )

 dense (Dense)               (None, 24, 8)             408

 flatten (Flatten)           (None, 192)               0

 dense_1 (Dense)             (None, 1)                 193

=================================================================
Total params: 2,556,401
Trainable params: 2,556,401
Non-trainable params: 0
_________________________________________________________________
```



```
string = ["Black people are a threat to the society"]
```

```
string_sequences = tokenizer.texts_to_sequences(string)
```

```
string_padded = pad_sequences(string_sequences, maxlen=max_length, padding="post", truncating="post")
```

```
predictions = model.predict(string_padded)
predictions
```

```
1/1 [==============================] - 0s 51ms/step

array([[0.9164899]], dtype=float32)
```

The trained model predicts the hate speech based on the given input text by giving a score between 0 to 1, the closer the value is to 1 the higher the hate speech.

REFERENCES

[1] T. A. Naidu and S. Kumar, "Hate Speech Detection Using Multi-Channel Convolutional Neural Network," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021

[2] A. Rodriguez, Y. -L. Chen and C. Argueta, "FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis," in IEEE Access, vol. 10, pp. 22400-22419, 2022

[3] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," in IEEE Access, vol. 9, pp. 109465-109477, 2021

[4] A. Kumar, V. Tyagi and S. Das, "Deep Learning for Hate Speech Detection in social media," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021

[5] F. T. Boishakhi, P. C. Shill and M. G. R. Alam, "Multi-modal Hate Speech Detection using Machine Learning," 2021 IEEE International Conference on Big Data (Big Data), 2021