Sentiment Analysis using Machine Learning

MINOR PROJECT I

Submitted by **Group 11**

Manas Tripathi (9920103196)
Lavish Arora (9920103193)
Pranjal Pateriya (9920103188)

Under the supervision of
Ms. Anuradha Gupta

Department of CSE/IT
Jaypee Institute of Information Technology University, Noida

DECEMBER 2022

ACKNOWLEDGMENT

We express our sincere gratitude to Ms. Anuradha Gupta, Department of CSE, Jaypee Institute of Information Technology, Noida for her invigorating guidance, continual encouragement, and supervision throughout the present work.

We also wish to extend our thanks to our batch mates for their insightful comments and constructive suggestions to improve the quality of this project work.

Credentials of students:

Pranjal Pateriya (9920103188)

Lavish Arora (9917103193)

Manas Tripathi (9917103196)

Signature of the mentor:

(Ms. Anuradha Gupta)

(Department of CSE&IT,

JIIT)

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Jaypee Institute Of Information Technology
Date: 28-11-2022

Pranjal Pateriya (9920103188)

Lavish Arora (9917103193)

Manas Tripathi (9917103196)

CERTIFICATE

This is to certify that the work titled "**Sentimental Analysis using Machine Learning**" submitted by Pranjal Pateriya:(9920103188), Lavish Arora:(9920103193) and Manas Tripathi(9920103196) of B.Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

ABSTRACT

Social media platforms are a huge platform to express opinions. We will use Machine Learning approach to determine multi-label emotions present in opinions expressed by people using various ML models to determine the user affiliation towards a political belief. We chose this topic to explore the field of Machine Learning. Analyzing Political affiliation of users gives an overview as to determine the direction of the political wind. One of the major motivations behind this topic was representation of data analysis in the documentary: The Great Hack (2019). “The Great Hack” blends in details about how the many strides in computer technology and data analysis now allow a massive, global expansion of a new type of social experiment, one that involves reshaping the world in a particular image. It unearths how data analysis from various platforms was used to if determine the behavior of a user and the masses in a particular area was against or under the influence of the Republican Party and then targeted the parts where they were close to a majority. This information was used for a variety of purposes meant to manipulate a certain cross-section of people. The master manipulators did not go after people whose minds had been made up; they went after on-the-fence folks referred to as “the persuadables.” Using the collected data, Cambridge Analytica set out to create fear and/or apathy to achieve the results of the political parties that hired them. Data Analysis and Algorithm was used to radicalization of the masses thus making it such a powerful tool.

Table of Contents

LIST OF FIGURES

ABBREVIATIONS

1. WORD2VEC Word to Vector
2. LSTM Long-short term memory
3. GRU Gated-Recurrent Units
4. RNN Recurrent Neural Networks
5. CNN Convolutional Neural Networks
6. Bi-LSTM Bi-Directional Long Term short memory

INTRODUCTION

Social media platforms are a huge platform to express opinions. We will use Machine Learning approach to determine multi-label emotions present in opinions expressed by people using various ML models to determine the user affiliation towards a political belief. We chose this topic to explore the field of Machine Learning. Analyzing Political affiliation of users gives an overview as to determine the direction of the political wind. One of the major motivations behind this topic was representation of data analysis in the documentary: The Great Hack (2019). “The Great Hack” blends in details about how the many strides in computer technology and data analysis now allow a massive, global expansion of a new type of social experiment, one that involves reshaping the world in a particular image. It unearths how data analysis from various platforms was used to if determine the behavior of a user and the masses in a particular area was against or under the influence of the Republican Party and then targeted the parts where they were close to a majority. This information was used for a variety of purposes meant to manipulate a certain cross-section of people. The master manipulators did not go after people whose minds had been made up; they went after on-the-fence folks referred to as “the persuadables.” Using the collected data, Cambridge Analytica set out to create fear and/or apathy to achieve the results of the political parties that hired them. Data Analysis and Algorithm was used to radicalization of the masses thus making it such a powerful tool.

Why it is important?

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools make that process quicker and easier than ever before, thanks to real-time monitoring capabilities.

The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organisations across the world.

The human language is complex. Teaching a machine to analyse the various grammatical nuances, cultural variations, slang and misspellings that occur in online mentions is a difficult process. Teaching a machine to understand how context can affect tone is even more difficult.

Therefore a sentiment analysis on various dataset is also used as a means to compare current prediction models and find their limitations and identifying gaps

.

BACKGROUND STUDY

RESEARCH PAPERS:

PAPER 1

Title of the Paper: Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods

Author: Iqra Ameer, Grigori Sidorov, Helena Gomez Ardorno, Rao Muhammad Adeel Nawab

Summary: Explores the need for a standard benchmark corpus to develop and evaluate multi-label classification methods. The multi-label emotion classification problem is not explored for code-mixed text, for example, English and Roman Urdu, although the code-mixed text is widely used in Facebook posts/comments, tweets, SMS messages, particularly by the South Asian community. For filling this gap, this study presents a large benchmark corpus for the multi-label emotion classification task, which comprises 11,914 code-mixed (English and Roman Urdu) SMS messages. Each code-mixed (English and Roman Urdu) SMS message manually annotated using a set of 12 emotions, including anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral (no emotion).

PAPER 2

Title of the Paper: Emotion Detection in Online Social Networks: A Multi-Label Learning Approach

Author: Xiao Zhang, Wenzhong Li, Haochao Ying, Feng Li, Siya Tang, Sanglu Lu

Summary: Emotion detection in online social networks (OSNs) can benefit kinds of applications, such as personalized advertisement services, recommendation systems, etc. Conventionally, emotion analysis mainly focuses on the sentence level polarity prediction or single emotion label classification, however, ignoring the fact that emotions might coexist from users' perspective.
To this end, in this work, they address the multiple emotions detection in OSNs from user-level view, and formulate this problem as a multilabel learning problem. First, it has discovered emotion labels correlations, social correlations, and temporal correlations from an annotated Twitter data set. Second, based on the above observations, it adopts a factor graph-based emotion recognition model to incorporate emotion labels correlations, social correlations, and temporal correlations into a general framework, and detect the multiple emotions based on the multilabel learning approach.

REQUIREMENT ANALYSIS

Software requirements

- Anaconda environment
- Jupyter Notebook
- TensorFlow
- Chrome browser

Hardware requirements:

- OS: Windows 7 or above (64-bit version)
- RAM: min 8 GB
- 2 GB min space
- Min 2 GB GPU

Language used:

- Python

DETAILED DESIGN

The methodology for analyzing public opinion incorporates (1) data collection, (2) pre-processing, (3)feature extraction (4) visualizes data.

Steps Involved:

Data collection:

One Pre-annotated dataset was available.

The second dataset was manually annotated.

Data Pre-Processing

After getting the dataset that contains tweets around the Indian Election, the next step was to clean the data to provide the input for the text classification model. Accuracy of feature extraction also greatly depends on the quality of text data. The following are the steps performed for data cleaning.

- **Normalization**

  This refers to the conversion of any non-text information into the textual equivalent. For this, we have used a library called normalize. This library is based on the nltk package, so it expects nltk word tokens.
- **Removal of Punctuation marks and symbols**

  A regular expression is used to eliminate the unnecessary punctuation marks(,;!'".?/* etc) and symbols like emojis, emoticons. removed. URLs, extra line feeds.
- **Tokenization and Removal of Stop Words**
- *Tokenize*:

  This breaks up the strings into a list of words or pieces based on a specified pattern usingRegular Expressions.
- *Stop Words*:

  Stop words are generally the most common words (such as "the", "a", "an", "in") in a language. These words are of no use because they don't help us to find the context or the true meaning of a sentence. We would not want these words to take up space in our database, or taking up the valuable processing time. These words were removed from the previously cleaned tweet text using a famous NLTK package nltk.stopwords.

WORD EMBEDDING

It's a representation of text where words that have the same meaning have a similar representation. In other words, it represents words in a coordinate system where related words, based on a corpus of relationships, are placed closer together. In the deep learning frameworks such as TensorFlow, Keras, this part is usually handled by an **embedding layer** which stores a lookup table to map the words represented by numeric indexes to their dense vector representations.

WORD2VEC

Gensim implementation of the word2vec model is used for the training of the dataset. The first step is to prepare the text corpus for learning the embedding by creating word tokens, removing punctuation, removing stop words, etc. This is done by the Tokenizer function of Keras. The word vector thus generated is passed to the gensim.models. Word2Vec function which trains the word embedding on the dataset. The model then uses this pre-trained word2vec embedding for the classification of tweets.

GLoVE

GloVe stands for global vectors for word representation. It is an unsupervised learning algorithm for generating word embeddings by aggregating a global word-word co-occurrence matrix from a corpus. The resulting embeddings show interesting linear substructures of the word in vector space.

FastText

FastText is an extension to Word2Vec proposed by Facebook in 2016. Instead of feeding individual words into the Neural Network, FastText breaks words into several n-grams (sub-words). For instance, the tri-grams for the word apple is app, ppl, and ple (ignoring the starting and ending of boundaries of words). The word embedding vector for apple will be the sum of all these n-grams. After training the Neural Network, we will have word embeddings for all the n-grams given the training dataset. Rare words can now be properly represented since it is highly likely that some of their n-grams also appears in other words.

IMPLEMENTATION

Classification Using Deep learning

Deep learning is a subset of machine learning in artificial intelligence (AI) that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network. Deep learning text classification model architectures generally consist of the following components connected in sequence:

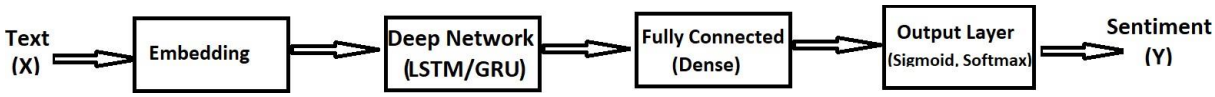


Fig 4.1 Deep learning architecture for text classification

Deep Learning architecture consists of layers with each layer's output working as an input to the next layer.As seen in the above diagram, the text is passed to the first layer which is the embedding layer, which thenpasses to the next layer and so on. To implement this architecture, we have used **Keras,** an open-source neural network library written in Python with a background working on TensorFlow.

Implementation

For text classification using deep learning architecture we have used the following layers:

Layer1: Embedding Layer
This layer makes use of pre-trained word embeddings (in our case, trained on our dataset).
The input to thislayer consists of the embedding matrix which was made using the trained word embedding on our dataset.

Layer2: Multilayer Perceptron:

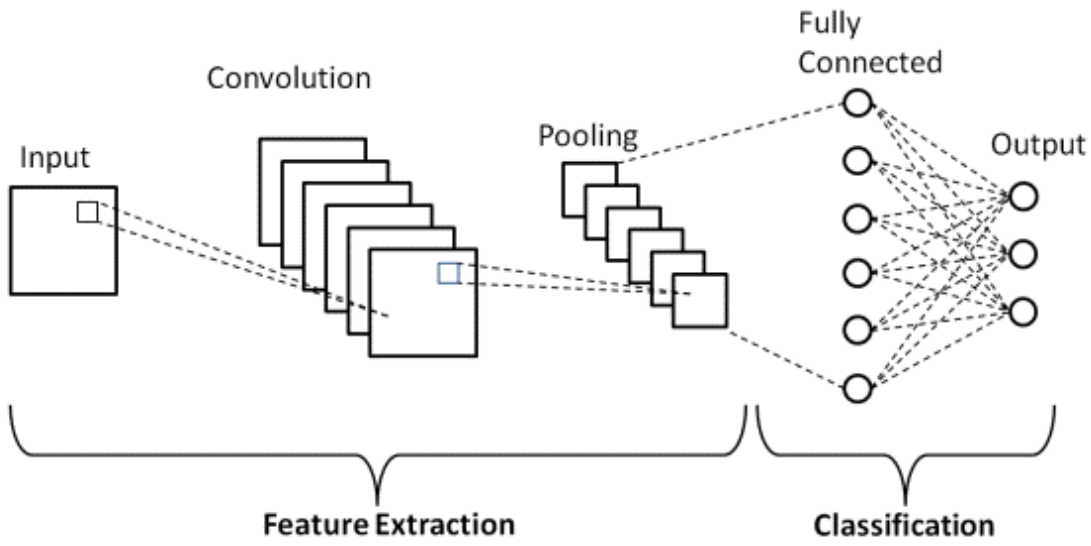
Sequential () imported from Keras acts as a hidden layer between the input and the output layer.

Long-short term memory and Bi-directional LSTM:

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell.

Concurrent neural network:

The first layers embed words into low-dimensional vectors. The next layer performs convolutions over theembedded word vectors using multiple filter sizes. For example, sliding over 3, 4, or 5 words at a time. Next, we max-pool the result of the convolutional layer into a long feature vector, add dropout regularization, and classify the result using a softmax layer.



Layer3: Output Layer

For binary classification, the activation function used is **'sigmoid'**. The number of output nodes depends on whether binary or multiclass classification is done.

After the layers are made, the compilation of the model is done. Compile choices used in our model:

- optimizer='adam'
- loss='binary_crossentropy'
- metrics='accuracy'

After compilation, the model was fit on a training data of 50% and validation and testing data of 25% each.

- Batch_size = 128
- Epochs = 10

Data Annotation:

A complete dataset of **1500** tweets were manually annotated and preprocessed.

1. Data Annotation for emotion category

Definitions of the emotions are as follows:

•**Anger:** The emotion anger, also known as annoyance or rage, is an extreme emotional condition. It includes an awkward and bitter response to an anticipated incitement, danger, or hurt.

•**Anticipation:** Anticipation is an emotion, including delight, excitement, or nervousness/anxiety because of or expecting an event, which also includes interest, hope, and prospect.

•**Disgust:** Disgust is a reaction to denial or refusal to something conceivably contagious, which also includes disinterest, loathing, and dislike.

•**Fear:** Fear is a feeling persuaded by an anticipated troublesome situation, or danger, which also includes anxiety, panic, and horror.

• **Joy:** Joy is a sensation of great pleasure, and happiness , which also includes ecstasy, pride, and delight.

•**Optimism:** Optimism is a mental attitude contemplating a trust or hope that the reaction of some particular aim, or conclusion in general, will be positive, supportive, and desirable , also includes confidence, certainty, and hopefulness.

•**Pessimism:** In general, pessimists likely to focus on the negatives of life, a depressed or negative mindset, also includes distrust, cynicism, and no confidence.

•**Surprise:** Surprise is a mental state that a person might feel if something unanticipated occurs, which also includes amazement and distraction.

•**Trust:** Usually refers to a circumstance defined by the following aspects: One group (trustor) is ready to depend on the activities of another group (trustee); the situation is directed to the future [31], also includes confidence, belief, and faith.

•**Neutral:** There is no emotion(s) in a sentence There are other emotions can be disappointment, shame, upset etc.

2.Data Annotation for opinion category

If given a tweet text show an opinion towards a any political party like BJP, Congress, BSP, AAP, INC, SP, RLD, CPI, PSP(L), AIMIM, JKNC, JD(S), RJD, etc. then it is marked as opinion.

3.Data Annotation for stance category

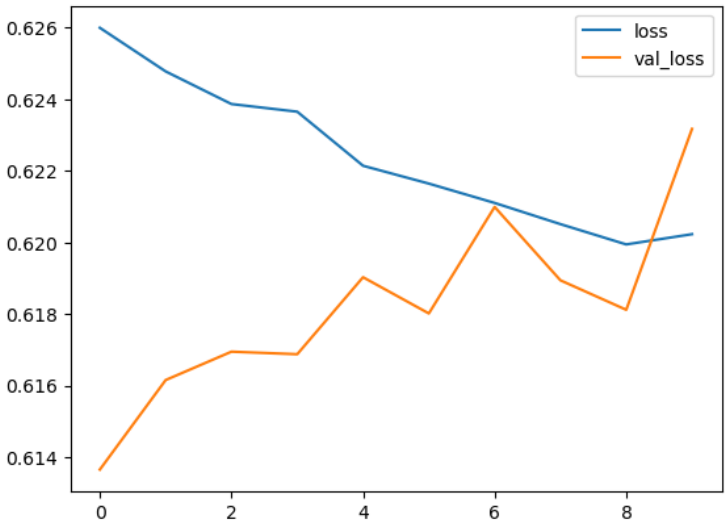
ProGovt -Given a tweet text show the user support towards BJP and its leadermark it is ProGovt.

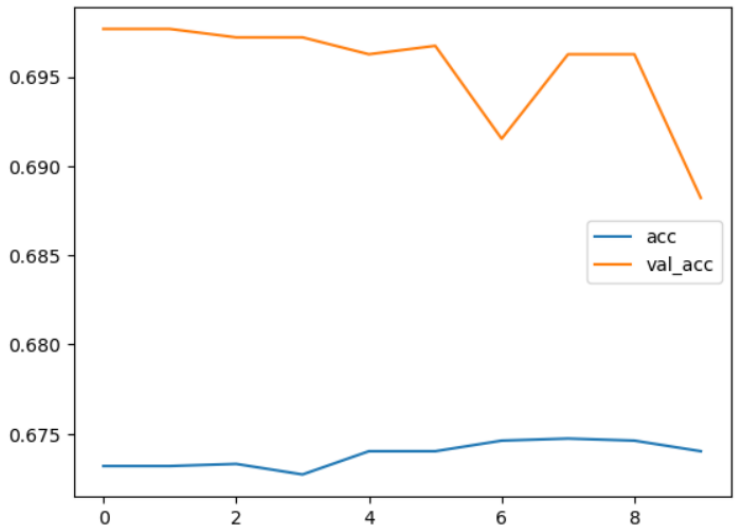
AntiGovt -Given a tweet text show that user is againsts BJP and its leader and show support to any other political party then it is marked as AntiGovt.

RESULTS AND ANALYSIS: Preprocessed-Dataset

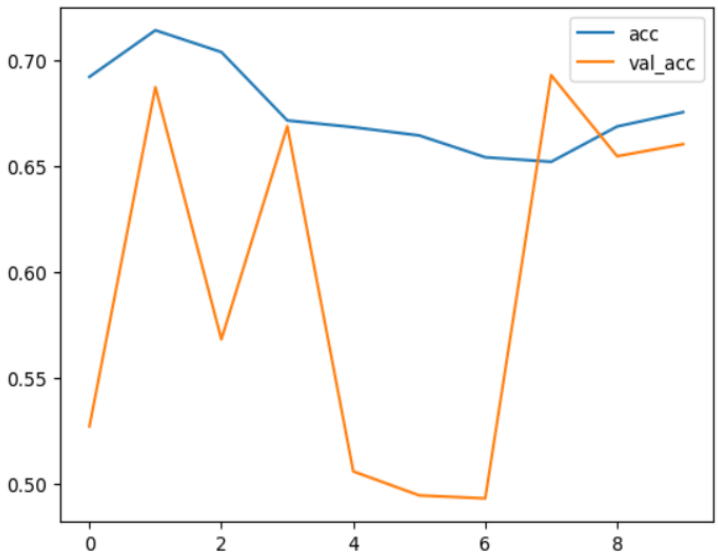
MODEL	FEATURE INPUT	ACCURACY	VALIDATION ACCURACY	LOSS	VALIDATION LOSS
LSTM	Word2Vec	67.40%	68%	62.02%	62.32%
CNN		67.57%	66%	39.03%	46.15%
Bi-LSTM		67.31%	60%	62.92%	61.26%

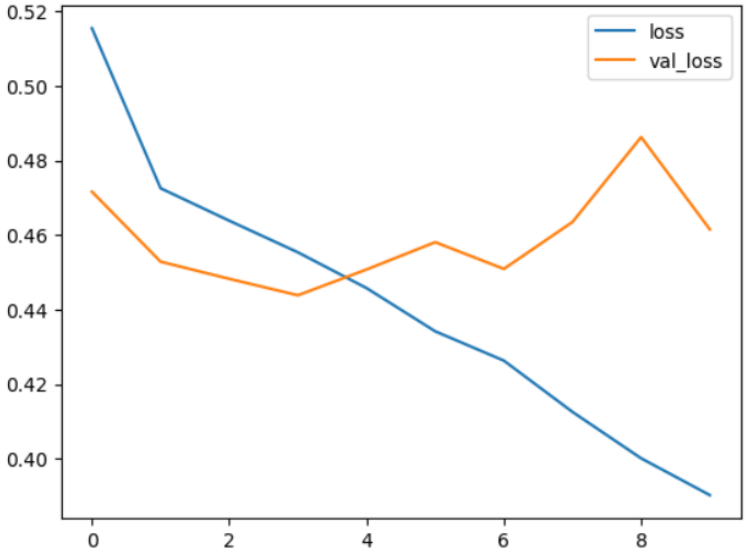
LSTM



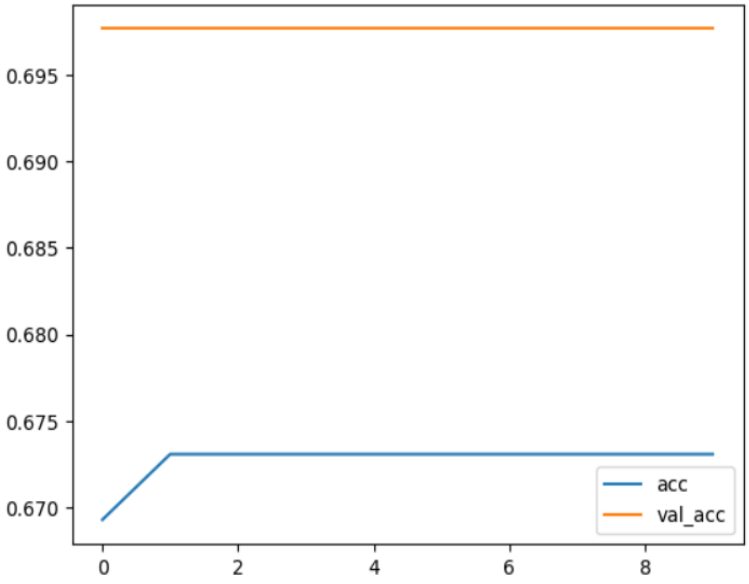


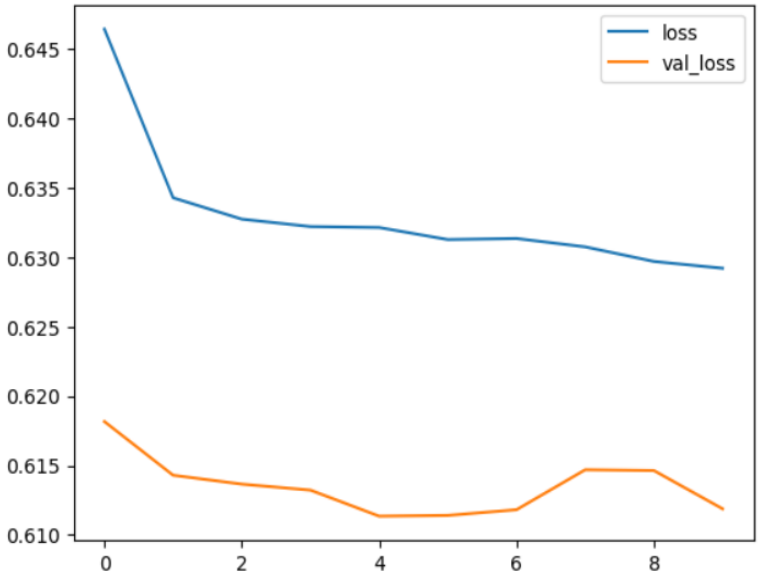
CNN





BI-LSTM



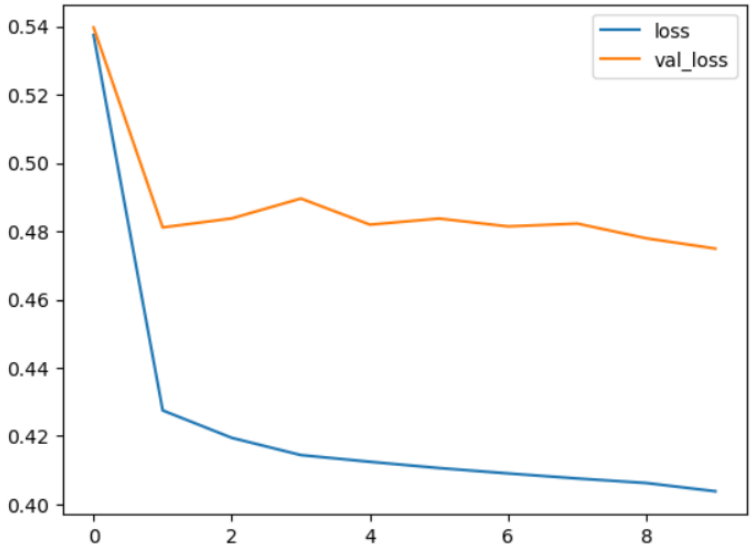


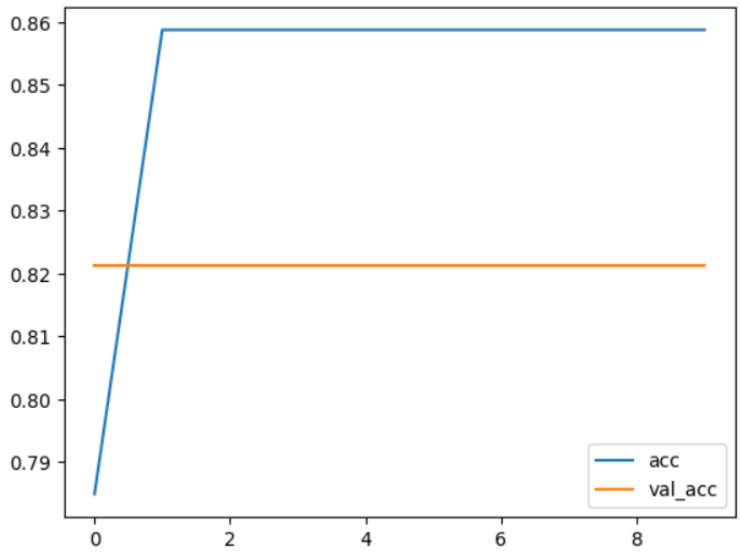
- CNN has performed slightly better than the other two models.

Annotated Dataset: 1500 tweets

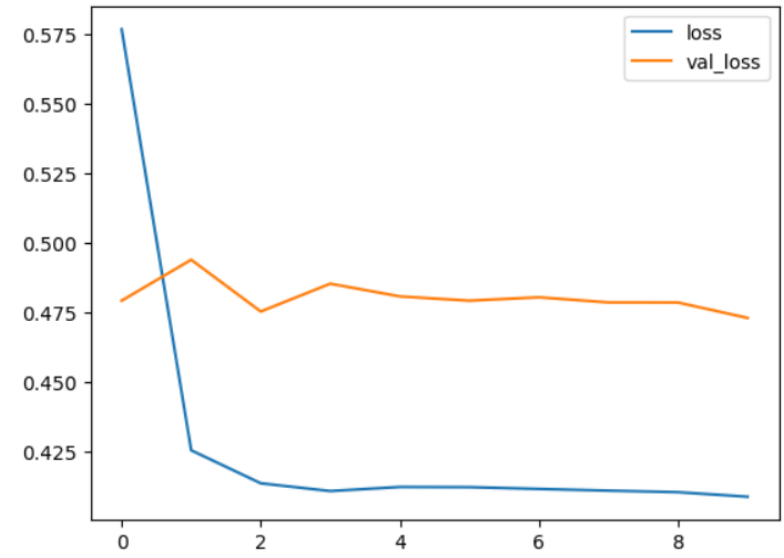
MODEL	FEATURE INPUT	ACCURACY	VALIDATION ACCURACY	LOSS	VALIDATION LOSS
LSTM	Word2Vec	85.87%	82%	40.38%	47.49%
CNN		61.43%	15%	1.16%	98%
Bi-LSTM		85.87%	82%	40.38%	47.49%

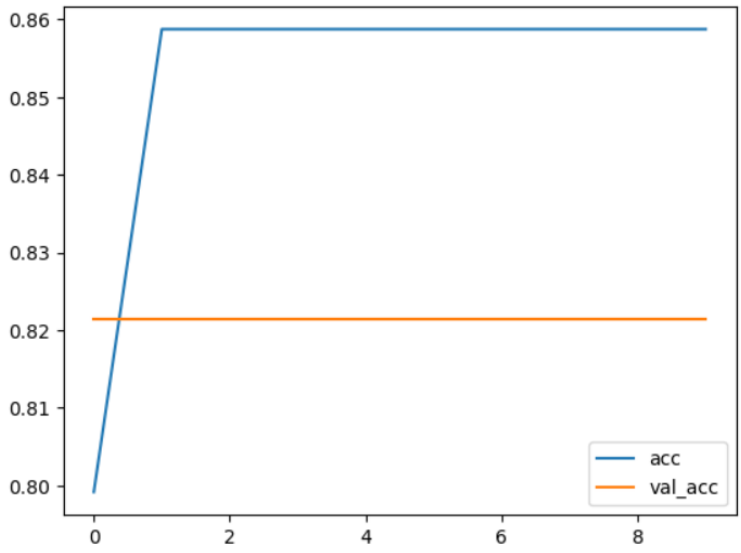
LSTM



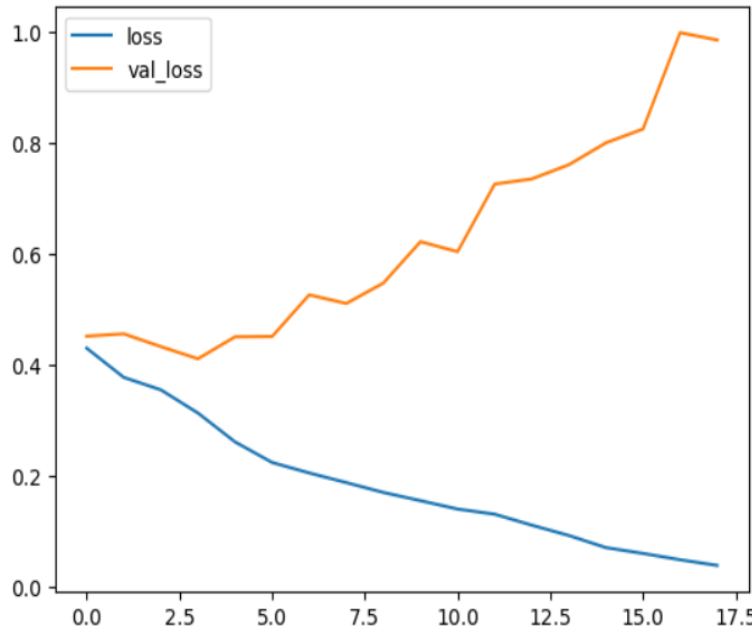


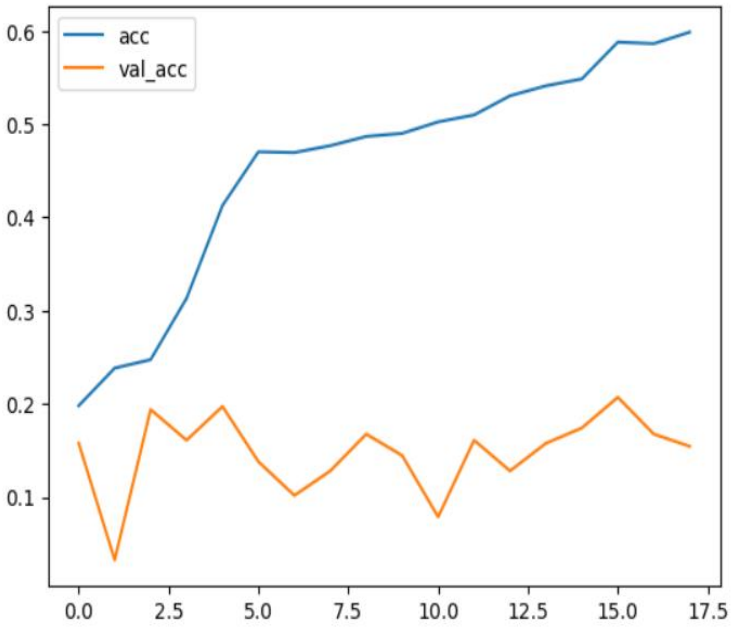
BI-LSTM





CNN





- LSTM and Bi-LSTM perform better than CNN.
- For low accuracy, underfitting can be one of the reasons which has led to bias in weights. Size is not enough

FUTURE SCOPE

The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens.

As a result of deeper and better understanding of the feelings, emotions and sentiments of a brand or organization's key, high-value audiences, members of these audiences will increasingly receive experiences and messages that are personalized and directly related to their wants and needs.

Again, sentiment analysis is on the verge of breaking into new areas of application. While we will likely always think of it first in terms of the traditional marketing sense, the world has already seen a few ways that sentiment analysis can be used in other areas.

CONCLUSION

Algorithms have long been at the foundation of most forms of analytics, including social media and sentiment analysis. With recent years bringing big leaps in machine learning and artificial intelligence, many analytics solutions are looking to these technologies to replace algorithms. Unfortunately for organizations looking to leverage sentiment analysis to measure audience emotions, machine learning isn't yet ready to tackle the complex nuances of text and how we talk, especially on social media channels that are rife with slang, sarcasm, double meanings and misspellings. These make it difficult for artificial intelligence systems to accurately sort and classify sentiments on social media. And, with any analysis project, accuracy is crucial. It is uncertain if machine learning will progress to the point that it *is* capable of accurately analyzing text, or if sentiment analysis projects will have to find a new basis to avoid the current plateau of algorithms.

References

- Multi-Label Emotion Classification on Code-Mixed Text: Data and Methods
- Emotion Detection in Online Social Networks: A Multi-Label Learning Approach
- https://medium.com/analytics-vidhya/understanding-embedding-layer-in-keras-bbe3ff1327ce#:~:text=Embedding%20layer%20is%20one%20of,word%20embeddings%20such%20as%20GloVe.
- https://analyticsindiamag.com/the-continuous-bag-of-words-cbow-model-in-nlp-hands-on-implementation-with-codes/
- https://github.com/lukasgarbas/nlp-text-emotion/blob/master/lstm_w2v_wiki.ipynb
- www.stackoverflow.com
- https://datascience.stackexchange.com/
- https://towardsdatascience.com/sentiment-analysis-using-lstm-and-glove-embeddings-99223a87fe8e
- www.youtube.com
- https://github.com/facebookresearch/fastText/tree/master/python