iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46694

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Study of Different Algorithms on Opinion Mining

Saurabh Dua[1], Ishita Sharma[2], Shaurya Singh[3], Kavisankar L[4]
[1, 2, 3]*B.Tech CSE, SRMIST, Kattankulathur, Tamil Nadu, India*
[4]*Assistant Professor, Department of CSE, SRMIST, Kattankulathur, Tamil Nadu, India*

Abstract: ***Social media is critical in today's world for exchanging information and disseminating ideas. A person's emotional impact has a significant impact on their day-to-day life. Sentiment analysis is a form of text mining that locates and pulls out subjective information from sources, allowing a company to track discussions online and monitor social sentiment about their brand, product, or service. Simply put, sentiment analysis helps determine the author's attitude towards a topic. Positive, neutral, or negative pieces of writing are classified by sentiment analysis software. Deep learning algorithms and various functions of natural language processing helps to interpret the written or spoken sentiments regarding a topic. An ecosystem where millions of bytes of data are produced daily has enabled sentiment analysis to be a key tool for interpreting these huge chunks of data. The purpose of this work is to conduct a sentiment analysis on "tweets" by making use of a variety of machine learning algorithms. The study will make an attempt to categorise the polarity of the tweet as either positive, negative, or neutral. In the event that a tweet has only positive, negative, or neutral components, the label assigned to the tweet will be determined by the sentiment that predominates.***

Keywords: Sentiment analysis, Convolutional Neural Networks, Long Short-Term Memory, Gated Recurrent Unit, Bidirectional Gated Recurrent Unit

I. INTRODUCTION

Internet platforms for social interaction such as Facebook, Twitter, LinkedIn and a slew of others have seen their user bases explode over the last decade[8]. Many of these folks are also being drawn into the conversation via social media through the use of hot-off-the-press insightful themes.

Recently, a lot of social media users have been using various social media platforms to express their views on a variety of topics. The popularity of these tweets has risen as a result of the vast number of individuals following them.

In addition, social media provides a wonderful chance platform for businesses to communicate with their customers quickly and effectively. Many people base their decisions on content created by others, such as comments on websites.

Prior to making a purchase, many people conduct research on a product. The promotion of a product and the dialogues occurring on social media sites like Facebook and Twitter are two important factors in a company's success[7].

To do a sentimental analysis based on the evaluations or remarks that people leave on social media. Using social media tags and a technique called Sentimental Analysis (SA), we can now determine whether or not the information we have presented is accurate[1].

It was crawled and classified positive/negative/neutral by Kaggle. To analyze and transform the data, the data comes with a variety of emoticons, usernames, and hashtags. As a way to represent a "Tweet," additionally, we are required to extract useful features such as bigrams and unigrams. The collected characteristics are fed into a variety of machine learning algorithms, which are then used to perform sentiment analysis. To get the best results, model ensembling is done instead of depending just on individual models. Finally, observations and conclusions from the experiments are done[4].

II. DATA DESCRIPTION

The data is given in the form of a comma-separated values file, and it contains tweets along with the emotions that are associated with them. In the dataset used for training, each tweet has a unique identity called tweet id, and the sentiment can be either positive, neutral or negative. Similarly, the dataset for testing is a CSV file of the type tweet id, tweet as its name suggests.

The dataset includes words, emoticons, symbols, URLs, and references to individuals. Emoticons and words help anticipate the mood, but references and URLs to persons don't help. Thus, URLs and citations can be ignored. In addition, the text has a variety of typos, excessive punctuation, and sentences with several repetitions of letters. To ensure that the dataset is consistent, the tweets must be pre-processed.

Both the practice and the test sets come with a total of 21465 and 5398 tweets respectively. The training dataset's shape is (21465,3), while the testing dataset's shape is (5398,2).

III. METHODOLOGY

The raw tweets that are scraped from Twitter are usually loud. As a result, people's use of social media is more casual than it used to be. The unique properties of tweets, such as retweets, emoticons, user mentions, etc., must be properly retrieved from the tweets[6]. To make it easier for classifiers to learn from raw Twitter data, it must be standardized. To standardize and condense the dataset, we've gone through a variety of pre-processing stages. A general pre-processing procedure for tweets is shown in Fig(I) :

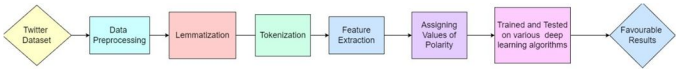


Fig (I)

IV. EXPERIMENTS

We test a wide range of classifiers in our research. As a precaution against overfitting, we only utilise 20% of the training dataset for validation, which is 17172 tweets for training and 4293 tweets for validation, respectively.

A. Long short-term memory (LSTM)

This deep learning method uses a neural network architecture that is based on synthetic recurrent networks (RNNs), commonly known as long short-term memory (LSTM). Feedforward neural networks are different from LSTMs. A LSTM network excels at detecting and predicting vital events in a particular statistic when there are gaps of uncertain duration. This particular problem can arise when conventional RNNs are trained in the presence of exploding and vanishing gradients. Because LSTMs are relatively insensitive to gap length, they are a better sequence learning approach than RNNs, hidden Markov models, and other sequence learning approaches. The model use two dropout layers, one dense and embedding layer which constitutes a total of 3,376,555 trainable parameters as shown in Fig(II).The probabilities of 'n' independent events are calculated by the SoftMax function. The probabilities of each target class across all target classes can be calculated by using this function.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 100)         3321800
_________________________________________________________________
lstm (LSTM)                  (None, None, 64)          42240
_________________________________________________________________
lstm_1 (LSTM)                (None, 32)                12416
_________________________________________________________________
dense (Dense)                (None, 3)                 99
=================================================================
Total params: 3,376,555
Trainable params: 3,376,555
Non-trainable params: 0
```

Fig (II)

B. Convolutional Neural Network (CNN)

Convolutional Neural Network, an artificial neural network is commonly employed in the fields of recognition and classification. To ensure the reliability of our forecasts, we have constructed a 7-layer model. Embedding layer 1 uses maximum features=100 and is the initial layer. This is a dropout layer, which is the second layer. The activation layer is set to relu, and there are 64 neurons in the convolutional layer's kernel size of 3. Layer 4 uses a 1D global maximum pooling algorithm. Dense layer 5 has 128 neurons and is activated by layer=relu. The dropout layer is the name for the sixth level. Dense layer 7 with activation layer=softmax. Categorical cross-entropy loss was employed in this model, with Adam as the optimizer and accuracy as the measure of choice as shown in Fig(III).

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 1133, 100)         3321800
_________________________________________________________________
dropout (Dropout)            (None, 1133, 100)         0
_________________________________________________________________
conv1d (Conv1D)              (None, 1133, 64)          19264
_________________________________________________________________
global_max_pooling1d (Global (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               8320
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 387
=================================================================
Total params: 3,349,771
Trainable params: 3,349,771
Non-trainable params: 0
```

Fig (III)

C. CNN + GRU

Kyunghyun Cho and colleagues used recurrent neural networks in 2014. The GRU is a recurrent neural network, similar to a long-short-term memory (LSTM) with a forget gate, but it lacks an output gate and so has fewer parameters. With this concept, the first layer is an embedding layer. In the second layer, we use a 1D convolutional kernel of size 3, activation=relu, and 64 neurons. The maximum area convolved is extracted using a Max Pooling layer with pool size=2. For the purposes of regularisation, the fourth layer is a dropout layer. In each iteration of the updates, 25% of the inputs will be dropped at random. The 128 neurons that make up the 5th layer form a gated recurrent unit. The sixth layer is a dropout layer, and its rate of 30% indicates that 30% of inputs will be dropped from consideration during each update iteration. The additional sixth layer is a smoothing layer. The seventh layer is the densest, consisting of 128 neurons with activation=relu. For this reason, the eighth layer is a dropout one with a rate of fifty percent. The ninth layer is a thick layer with activation of softmax and an output of 3. Finally, an accurate measure was selected, the adam optimizer was used, and categorical cross entropy loss was applied to compile the models. All the parameters were shown in Fig (IV).

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_2 (Embedding)      (None, 1133, 100)         3321800
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 1133, 64)          19264
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 566, 64)           0
_________________________________________________________________
dropout_2 (Dropout)          (None, 566, 64)           0
_________________________________________________________________
gru (GRU)                    (None, 566, 128)          74496
_________________________________________________________________
dropout_3 (Dropout)          (None, 566, 128)          0
_________________________________________________________________
flatten (Flatten)            (None, 72448)             0
_________________________________________________________________
dense_3 (Dense)              (None, 128)               9273472
_________________________________________________________________
dropout_4 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_4 (Dense)              (None, 3)                 387
=================================================================
Total params: 12,689,419
Trainable params: 12,689,419
Non-trainable params: 0
```

Fig (IV)

D. Bidirectional GRU

A BiGRU, or Bidirectional GRU, is a sequence processing model that employs not one but two GRUs. one that processes the information forward and one that processes it backward. The only gates in this one-way recurrent neural network are the input and forget ones. After an embedding layer comes a 25% spatial dropout layer. A 128-bit size bidirectional GRU layer constitutes the third layer. It's important to note that the fourth layer is a 50% dropout rate layer. Activation = softmax is used in the fifth dense layer, which has an output of 3. Compilation of the model was done using the Adam optimizer and category cross entropy as shown in the Fig(V).

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 1133, 100)         3321800
_________________________________________________________________
spatial_dropout1d (SpatialDr (None, 1133, 100)         0
_________________________________________________________________
bidirectional (Bidirectional (None, 256)               176640
_________________________________________________________________
dropout_5 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_5 (Dense)              (None, 3)                 771
=================================================================
Total params: 3,499,211
Trainable params: 3,499,211
Non-trainable params: 0
```

Fig (V)

V. RESULTS

With the median F1 score, the model was able to classify tweets as positive, negative, or neutral. These findings are the outcome as shown in Fig(VI):

Methods	F1-Score	Advantages	Disadvantages
LSTM	0.64126	To efficiently manage long-term dependencies, LSTM relies on its ability to store information temporarily.	Long short-term memory devices are delicate to initializations with random weights.
CNN	0.65446	As a result of its simplified structure and reduced number of parameters, Convolutional Neural Networks (CNNs) are more manageable throughout the learning process.	If the CNN is composed of several layers, the training procedure will take a significant amount of time[8].
CNN+ GRU	0.64451	The advantage of CNN over RNN is that, instead of having to name each individual node, we just have to label entire phrases.	The emotional polarity we get from a statement is the same regardless of whether or not it contains aspect information, which it is unable to represent.
Bidirectional GRU	0.62343	Data may be processed in both directions using a bidirectional GRU, and the output layer is comprised of data from the two independent hidden layers.	It can be challenging to train because it accepts input both forward and backward..

Fig (VI)

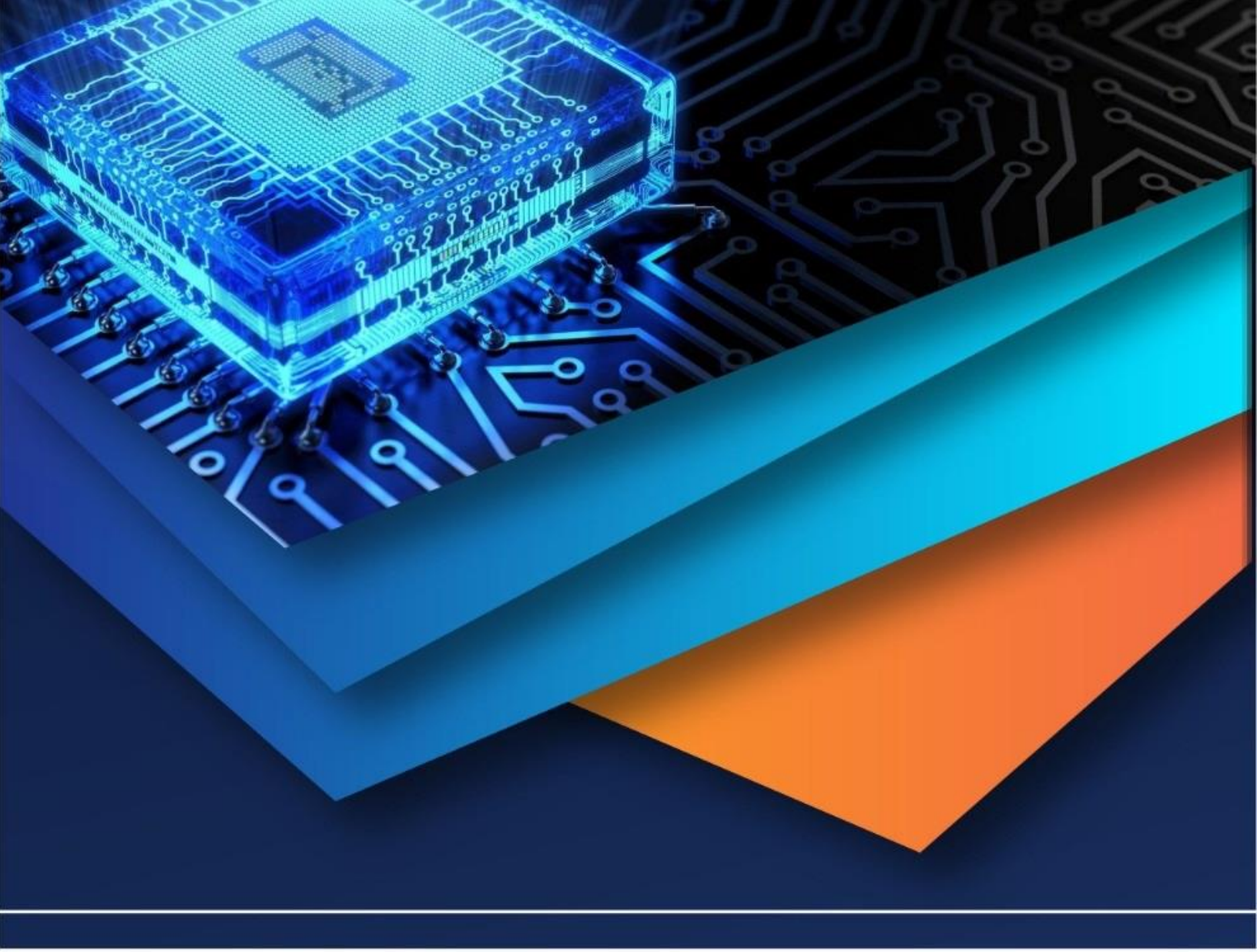
VI. CONCLUSION

In this study, we compare four distinct approaches to sentiment analysis to see which one yields the most reliable results. The CNN model is determined to have the highest F1 score (0.65446) of all the models tested.

When it comes to handling emotional ranges and using symbols, there is a lot of room for development. Our models may be trained to manage a wide spectrum of emotions. Similarly, positive and negative sentiments can have varying degrees of positivity, as seen by the sentences "This is good" and "This is outstanding," respectively. The sentiment can be divided into a number of categories, such as the range from -2 to +2.

REFERENCES

[1] Ahmed Hassan Yousef, Walaa Medhat and Hoda K. Mohamed, Nile University, "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, Volume 5, Issue 4, May 2014.
[2] Mohamed Hayouni and Sahbi Baccar, ESIGELEC, "Sentiment Analysis Using Machine Learning Algorithms", August 2021.
[3] Md. Serajus Salekin Khan, Sanjida Reza Rafa, Al Ekram Hossain Abir and Amit Kumar Das, East West University (Bangladesh), "Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity" Vol. 2 No. 03, July 2021.
[4] Sunmoo Yoon, Faith E Parsons, Kevin Joseph Sundquist and Jacob Julian, Columbia University "Comparison of Different Algorithms for Sentiment Analysis: Psychological Stress Notes", Stud Health Technol Inform. Author manuscript, January 2017
[5] Nikhil George, Tinto Anto and Niranjan Rao, "A Case Study on the Different Algorithms used for Sentiment Analysis", International Journal of Computer Applications, Volume 138 – No.12, March 2016
[6] Paolo Fornacciari, Monica Mordonini and Michele Tomaiuolo, Università di Parma, "A Case-Study for Sentiment Analysis on Twitter", January 2018
[7] Soumi Sarkar, National Institute of Technology, Durgapur, "Sentiment Analysis in Twitter: A Case Study in the Indian Airline Industry", International Journal Of Data Mining And Emerging Technologies, Volume: 7, Issue: 2, January 2017
[8] Balaji Karumanchi, "An Unsupervised Clustering Approach for Twitter Sentimental Analysis: A Case Study for George Floyd Incident", International Journal of Computer Trends and Technology (IJCTT), Volume-68 Issue-6, June 2020
[9] Sarah Shukri, Rawan I. Yaghi, Ibrahim Aljarah and Hamad Alsawalqah, University of Jordan, "Twitter Sentiment Analysis: A Case Study in the Automotive Industry", 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), November 2015
[10] Vartika, C. Rama Krishna, Ravinder Kumar and Yogita, "Sentiment Analysis of Train Derailment in India: A Case Study from Twitter Data" 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), September 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY