AN INDEPTH ANALYSIS OF CATEGORIZED MINING ALGORITHMS FOR OPINION MINING[1]

Shubham Bhardwaj

National Institute of Technology, Hamirpur, Himachal Pradesh

DOI: 10.37648/ijrst.v10i01.011

Received: 10th January, 2020; Accepted: 05th March, 2020; Published: 29th March, 2020

ABSTRACT

Today's information and ideas can't be shared without social media. A person's day-to-day life is significantly affected by their emotional impact. An ecosystem that generates millions of bytes of data daily makes sentiment analysis essential for interpreting these enormous amounts of data. Sentiment analysis, a type of text mining, finds and extracts personal information from various sources, allowing businesses to monitor social sentiment about their brand, product, or service. Simply put, sentiment analysis enables one to ascertain the author's perspective on a topic. Writing is categorized as either positive, neutral, or negative by the software for sentiment analysis. With the help of deep learning algorithms and natural language processing functions, written or spoken sentiments about a topic can be better understood.

This work uses various machine learning algorithms to conduct sentiment analysis on "tweets." The predominant sentiment will determine the label given to a tweet if the tweet only contains positive, negative, or neutral elements. The study will attempt to classify the tweet's polarity as positive, negative, or neutral.

INTRODUCTION

Over the past ten years, the number of people using social networking websites like Facebook, Twitter, and LinkedIn has increased dramatically [8]. Many of these people are also being drawn into the conversation via social media through fresh, insightful topics.

Numerous social media users have recently used a variety of social media platforms to voice their opinions on various subjects. The large number of people who follow these tweets has increased their popularity.

Additionally, businesses can communicate with their clients efficiently and quickly thanks to social media. Many people make decisions based on other people's content, like comments on websites.

Many people research a product before making a purchase. Two important aspects of a company's success are product promotion and social media conversations like those on Facebook and Twitter [7].

To do a nostalgic investigation given the assessments or comments that individuals leave via virtual entertainment. Using social media tags and a technique known as Sentimental Analysis (SA)[1], we can now determine whether our presented information is accurate. Kaggle crept in and categorized it as positive, negative, or impartial. Emoticons, usernames, and hashtags that can be

[1] *How to cite the article:* Bhardwaj S., An In-depth Analysis of Categorized Mining Algorithms for Opinion Mining, IJRST, Jan-Mar 2020, Vol 10, Issue 1, 53-57, DOI: http://doi.org/10.37648/ijrst.v10i01.011

used for analysis and transformation are included in the data. To represent a "Tweet," we must extract useful features like bigrams and unigrams. Sentiment analysis is carried out by feeding the collected characteristics into various machine learning algorithms. Instead of relying on individual models, model ensembling is used to achieve the best results. The experiments' findings and conclusions are finally completed [4].

DESCRIPTION OF DATA

The tweets and their feelings are included in the data, provided in a values file separated by commas. The sentiment of each tweet in the training dataset can be either positive, neutral, or negative. Each tweet has a unique identifier known as a tweet id. Similarly, the dataset that will be tested is a CSV file with the name "tweet id, tweet." The dataset incorporates words, emojis, images, URLs, and references to people. Emoticons and words help predict the mood, but it's important to clarify people's names and URLs. As a result, citations and URLs can be ignored. Furthermore, the text has an assortment of grammatical errors, excessive accentuation, and sentences with a few reiterations of letters. The tweets need to be pre-processed to guarantee consistency throughout the dataset.

METHODOLOGY

The scraped tweets from Twitter are typically loud. As a result, social media use has become less formal than it once was. Tweets' distinctive features, such as retweets, emoticons, and user mentions, must be retrieved correctly from tweets [6].

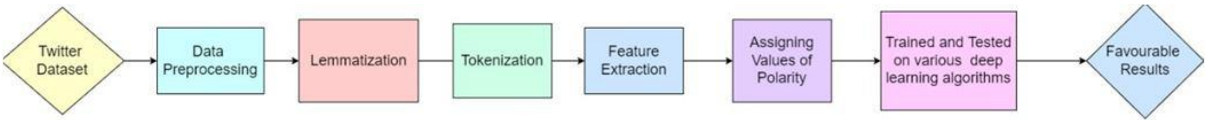


Figure 1

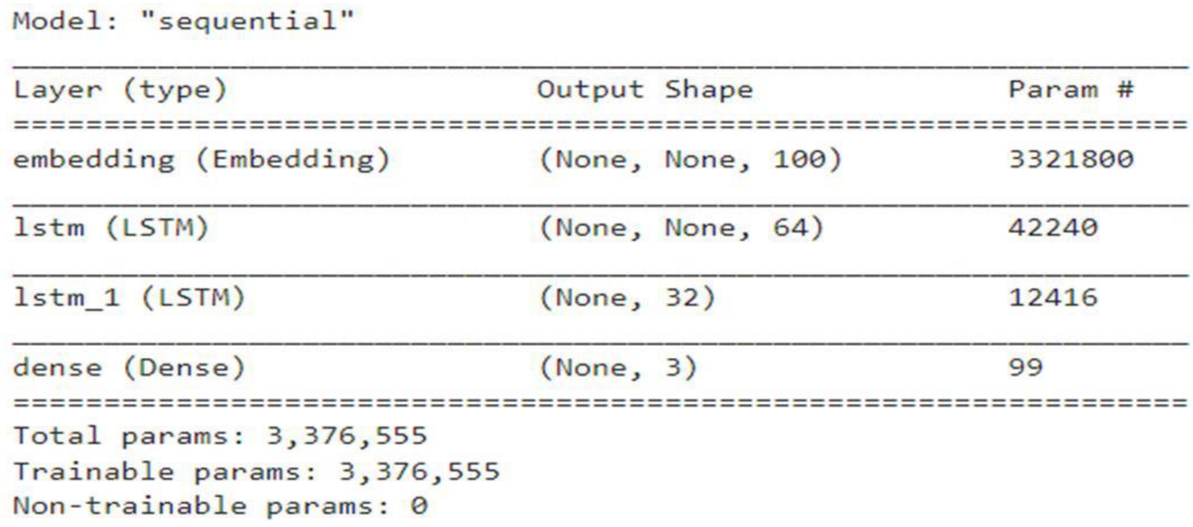
It needs to be standardized to make it simpler for classifiers to learn from raw Twitter data. To standardize and simplify the dataset, we went through several pre-processing steps. Figure (I) depicts a typical tweet pre-processing process:

EXPERIMENTS

In our research, we test a wide variety of classifiers. We use only 20% of the training dataset for validation to avoid overfitting, using 17172 tweets for training and 4293 tweets for validation, respectively.

A. LSTM

A neural network architecture based on RNNs, more commonly referred to as LSTM, is used in this deep learning strategy. LSTMs and feedforward neural networks are not the same things. At the point when there are holes of uncertain terms, an LSTM network succeeds at distinguishing and foreseeing critical occasions in a specific measurement. When conventional RNNs are trained with gradients that explode and disappear, this problem may occur. Because they are less sensitive to gap length than RNNs, hidden Markov models, and other sequence learning techniques, LSTMs are a superior sequence learning strategy. The model has two dropout layers—one dense and one embedding layer—with 3,376,555 trainable parameters, as shown in Figure (II). The SoftMax function calculates the probabilities of 'n' independent events. The probabilities of each target class across all target classes can be calculated with this function.

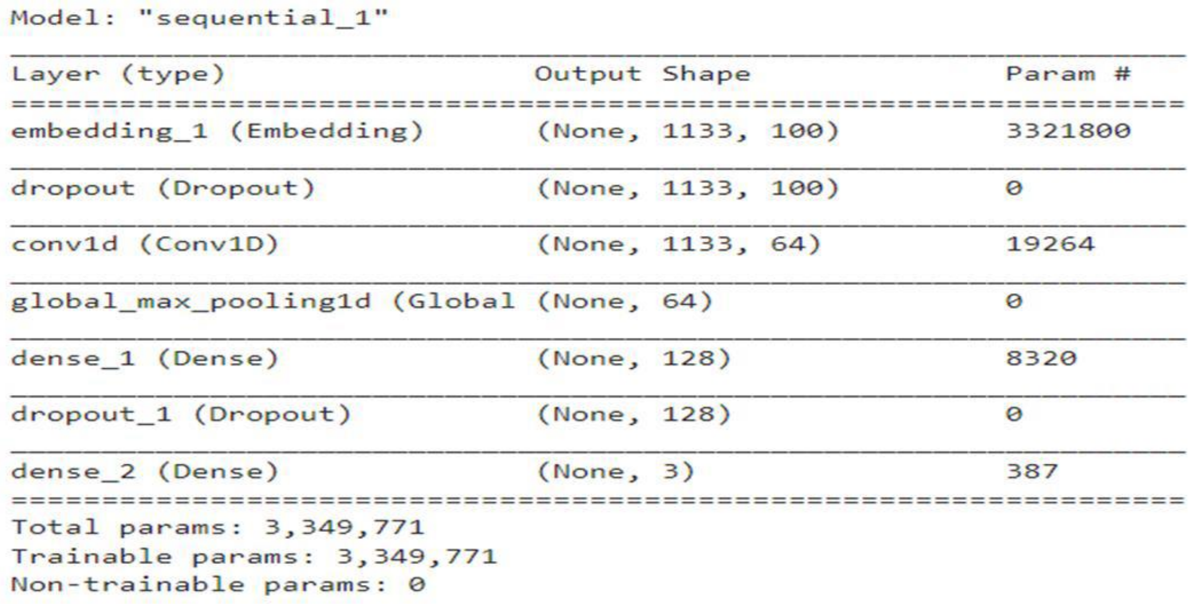


```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 100)         3321800
_________________________________________________________________
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
```

Figure 2

B. CNN

Recognition and classification are two common applications for the convolutional neural network, an artificial neural network. We have developed a seven-layer model to guarantee the accuracy of our forecasts. The initial layer, embedded layer 1, has a maximum feature count of 100. A dropout layer like this makes up the second layer. The convolutional layer has 64 neurons in its three-kernel size, and the activation layer is set to relu. Layer 4 employs a 1D global maximum pooling algorithm. Layer=relu triggers the firing of 128 neurons in dense layer 5. The dropout layer is the name of the sixth level. The softmax activation layer and dense layer 7 This model used category cross-entropy loss, Adam as the optimizer, and accuracy as the preferred measure, as shown in Figure (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
```

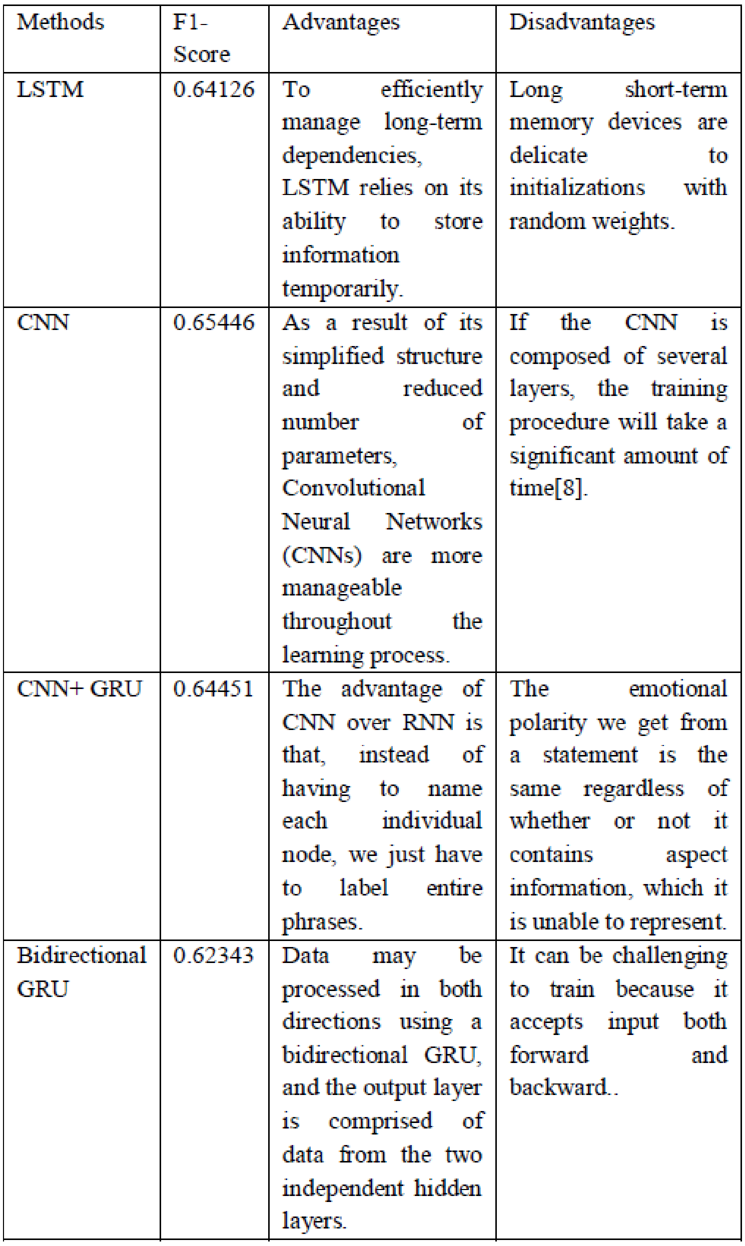
Figure 3

C. BGRU

A bidirectional GRU, or BiGRU, is a sequence processing model that uses not just one but two GRUs.one that process information in both the forward and reverse directions. The input and forget are the only gates in this one-way recurrent neural network. A 25% spatial dropout layer follows an embedding layer. The third layer is a bidirectional GRU layer of 128 bits. It is essential to keep in mind that the fourth layer has a 50% dropout rate. In the fifth dense layer, activation = softmax is used with an output of 3. The Adam optimizer and category cross-entropy were used to compile the model, as depicted in Fig(V).

RESULTS

Based on the median F1 score, the model could classify tweets as positive, negative, or neutral. The result is represented in Fig. VI by these findings:



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..

Figure 4

CONCLUSION

In this review, four particular ways to deal with feeling examination are contrasted with figure out which delivers the most solid outcomes. Out of all the models that were tested, the CNN model had the highest F1 score, which was 0.65446.

The management of dynamic ranges and the use of symbols both require significant improvement. It's possible that our models are prepared to deal with a variety of emotions. Positive and negative emotions can have varying degrees of positivity, similar to the phrases "This is good" and "This is outstanding. "From -2 to +2, there are several sentimental categories.

REFERENCES

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.

Mohamed Hayouni and Sahbi Baccar, ESIGELEC, “Sentiment Analysis Using Machine Learning Algorithms”, August 2021.

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.

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 Technology

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

Paolo Fornacciari, Monica Mordonini and Michele Tomaiuolo, Università di Parma, “A Case-Study for Sentiment Analysis on Twitter”, January 2018

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

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

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

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.