Sentiment Analysis on IMDB dataset - Movie Review

PG 34 Yuvraj S. Chauhan, PG 37 Naveen Maheshwari, PG 36 Rashmita Chauhan, PG 44 Shubham Shail

Mentored by Prof. Vaishali Suryawanshi

Computer Science Engineering, World Peace University
Kothrud, Maharashtra India

***Abstract*—** With the growing amount of data in reviews, it is quite prudent to automate the process, saving on time. Sentiment analysis is an important field of study in machine learning that focuses on extracting information of subject from the textual reviews. The area of analysis of sentiments is related closely to natural language processing and text mining. It can successfully be used to determine the attitude of the reviewer in regard to various topics or the overall polarity of the review. In the case of movie reviews, along with giving a rating in numeric to a movie, they can enlighten us on the favorableness or the opposite of a movie quantitatively; a collection of those then gives us a comprehensive qualitative insight on different facets of the movie. Opinion mining from movie reviews can be challenging due to the fact that human language is rather complex, leading to situations where a positive word has a negative connotation and vice versa. In this study, the task of opinion mining from movie reviews has been achieved with the use of neural networks trained on the " IMDB Movie Review Database"

i. INTRODUCTION

The main objective of this project is to develop a sentiment analysis system by classifying the sentiment of the sentence from the IMDB reviews and show emoticons to the sentence ,we will be trying different machine learning and deep learning approaches to classify the sentiments ,we will also be using word tokenization and identifying different stop words from the dataset to stem the data and we have used neural networks for data classification ,there are three algorithms that we will be trying which are naïve Bayes approach, Recurrent neural network and artificial neural network and we will be showing the accuracy of each approach in our project.,

LITERATURE REVIEW

S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, 2020, pp. 1-4, doi: 10.1109/ICCIS49240.2020.9257657.

In addition, we do the text processing from data obtained and use Naive Bayes method to predict the class. Afterward, compare with other methods such as SVM and KNN. We classify by two classes namely positive and negative.

From the results of our experiments, it can be seen that the Naïve Bayes method has a better accuracy level (i.e. 80.90%) compared to using other methods, such as KNN which only has an accuracy rate of 75.58% and an accuracy rate using SVM which is 63.99%.

Wongkar, Meylan & Angdresey, Apriandy. (2019). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. 1-5. 10.1109/ICIC47613.2019.8985884.

In this paper, a Long Short-Term Memory classifier is used with Adam optimizer to automatically categorize the preprocessed IMDb movie reviews. In total 10k reviews are considered, 5k for positive and 5k for negative sentiments.

Results have concluded that the highest accuracy attained by the devised approach is of 89.9%.

A superior accuracy can be attained by using further data preconditioning techniques. Furthermore, higher classification accuracy can be achieved by employing the ensemble classifiers or deep learning approaches. Exploring these opportunities is another prospect.

Performance Analysis of Different Neural Networks for Sentiment Analysis on IMDb Movie Reviews
Md. Rakibul Haque;Salma Akter Lima;Sadia Zaman Mishu
2019 3rd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)

In this paper, we have compared between CNN, LSTM and LSTM-CNN architectures for sentiment classification on the IMDb movie reviews in order to find the best-suited architecture for the dataset.

Experimental results have shown that CNN has achieved an F-Score of 91% which has outperformed LSTM, LSTM-CNN and other state-of-the-art approaches for sentiment classification on IMDb movie reviews.

Albert-based sentiment analysis of movie review
Zhongxiang Ding;Yali Qi;Deping Lin
2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)

In order to improve the user experience and ease the endurance of mobile terminals, this paper studies the task offloading strategy for optimizing the energy consumption of mobile terminals.

At present, most of the research only focuses on a single cloud offload and edge offload, and the research on reducing the energy consumption of mobile terminals is insufficient.

IMDb Sentiment Analysis COMP 551 - Group 17
Authors : Beatrice Lopez ,Minh Anh Nguyen and Xavier Sumba February 2019

In this research they have proposed model for sentiment analysis using RNN and word2vec. RNN in this model is implemented using framework Tensorflow. The research results show that the model approach has better accuracy with other machine learning models with result of accuracy is 91.98%.

Kurniasari, Lilis & Setyanto, Arif. (2020). Sentiment Analysis using Recurrent Neural Network. Journal of Physics:

Conference Series. 1471. 012018.
10.1088/1742-6596/1471/1/012018.

From this research, they also have to pay attention to the possibility of overfitting the model when carrying out the testing process. In the future They can try to use RNN and LSTM to overcome overfitting problems and to improve the performance of the model.

The working principle of deep learning using a neural network follows the architecture of neural networks in the human brain. One of the neural network architectures used in analytic sentiment is recurrent neural networks (RNN). RNN algorithm will associate each word in the input with a certain time step.

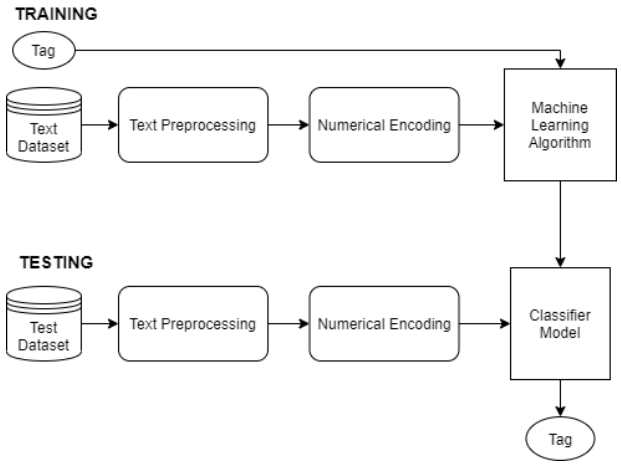
The proposed hybrid model is curated in three steps.

The first step is the preprocessing stage where impurities correction and multi aspect based filtration is applied. The spell correction, stemming,

abbreviation expansion, stop words removal is defined in this stage to normalize the input tweets.

The second stage takes the filtered text for processing in order to define statistical features. The initial training and testing sets are converted into

feature sets.



The features obtained in the second step are evaluated via the use of the hybrid classifier. Sentiment prediction is the final stage of this model. In

classification stage, the probabilistic predictive decision is applied for selection of KNN or SVM classifier for individual instance.

‘

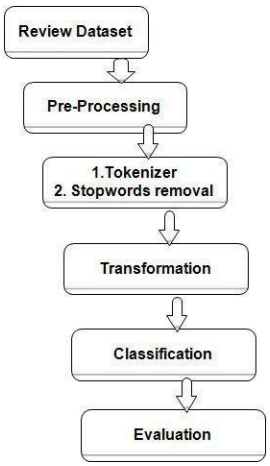
ii. PROPOSED METHODOLOGY

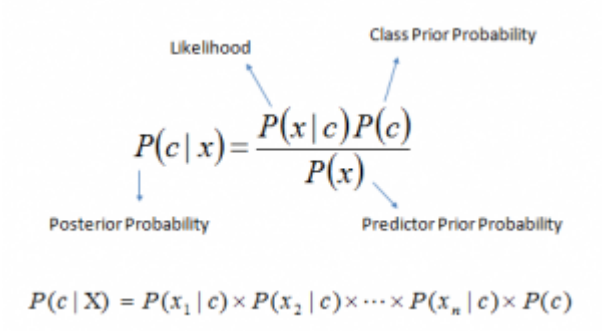
There are different architecture in our proposed methodology that we are following.

Naive Bayes

We use this architecture in natural language processing due to its ability that considers every part of the sentence as an equal contribution to the results and every word is independent of each other.

Fig. Analysis Architecture :





https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

In NLP we perform all the operations of NLP pipelines then calculate results using

P(positive | overall liked the movie) *= P(overall liked the movie | positive) * P(positive) / P(overall liked the movie)*

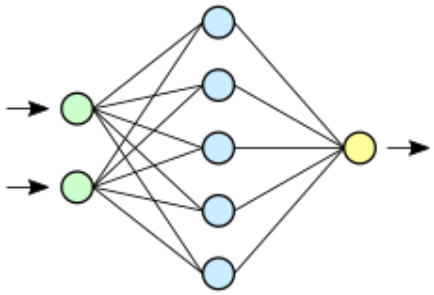
Artificial Neural Network

We have also tried to develop an architecture using an artificial neural network.

https://i2.wp.com/techvidvan.com/tutorials/wp-content/uploads/sites/2/2020/05/Architecture.jpg?ssl=1

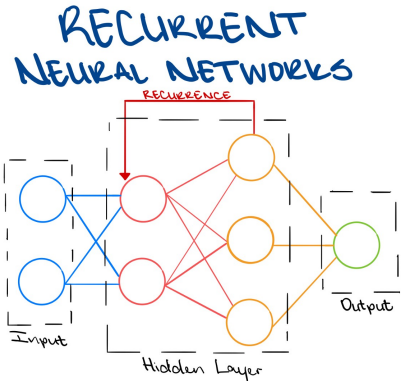
In this, all the sentences are preprocessed and then put through the inputs.
After all, this has happened then activation functions provide values to the different layers of architecture which eventually gives you the output which is binary.

In this, we can use ANN to find the sentiment due to its ability to perform well during binary classification**.**



Recurrent Neural Network

The problem with all the other architectures is that their mainframe of finding the sentiment was not with memory retained. Here we use the longest short-term memory to find the sentiment.



After all the preprocessing of sentences has been done then with LSTM and proper sequencing with a post or pre padding we get the results of the model here which is binary

Results

For Naive Bayes

```
[[4302  733]
 [1646 3319]]
0.7621
```

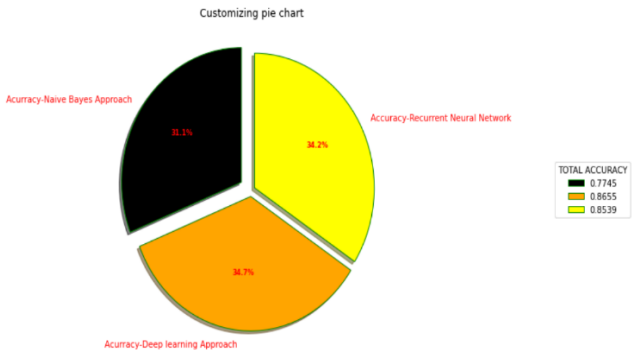
FOR ANN

```
[[4130  835]
 [ 790 4245]]
0.8375
TIME REQUIRED FOR THE JOURNEY
1277.9821286201477
```

FOR RNN

```
Correct Prediction: 8540
Wrong Prediction: 1460
Accuracy: 85.39999999999999
```

Visualization



iii. Results and Discussion

The system was tested by us ,we used gradio interface for testing in which we got the visual representation of the classified texts been passed in the form of emotes..As we can see that two different texts have been passed and our system can detect whether the sentence is positive or negative ,false positive or false negative and an emote is displayed as follows.

.

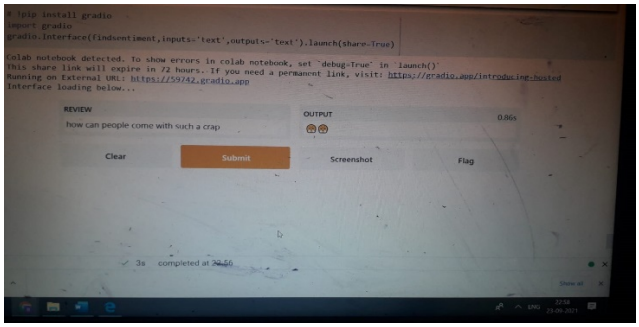


FIG 1

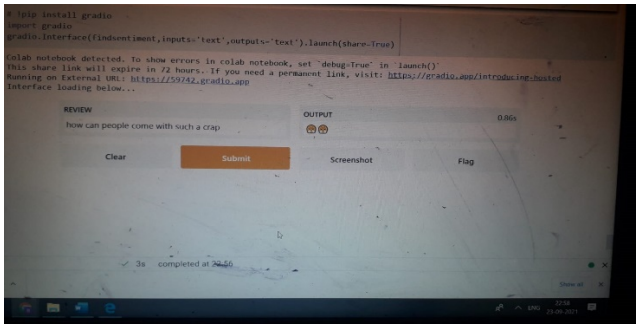


Fig 2

iv. Dataset and Features

IMDB DATASET HAVING 50K MOVIE REVIEWS FOR NATURAL LANGUAGE PROCESSING OR TEXT ANALYTICS.
THIS IS A DATASET FOR BINARY SENTIMENT CLASSIFICATION CONTAINING SUBSTANTIALLY MORE DATA THAN PREVIOUS BENCHMARK DATASETS. WE PROVIDE A SET OF 25,000 HIGHLY POLAR MOVIE REVIEWS FOR TRAINING AND 25,000 FOR TESTING. SO, PREDICT THE NUMBER OF POSITIVE AND NEGATIVE REVIEWS USING EITHER CLASSIFICATION OR DEEP LEARNING ALGORITHMS.

PUBLICATIONS USING THE DATA SET:
Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).*

v. Comparative Analysis

Based on the test we performed on the system we could find our system to be 85.399% efficient using Recurrent neural network.algorithm.

```
[ ]  # y_pred=model.predict(x_test)
     # y_pred=np.argmax(y_pred,axis=1)
     y_pred = (model.predict(x_test) > 0.5).astype("int32")


     # y_pred = model.predict_classes(x_test, batch_size = 128)
     true = 0
     for i, y in enumerate(y_test):
       if y == y_pred[i]:
         true += 1
     print('Correct Prediction: {}'.format(true))
     print('Wrong Prediction: {}'.format(len(y_pred) - true))
     print('Accuracy: {}'.format(true/len(y_pred)*100))

     Correct Prediction: 8540
     Wrong Prediction: 1460
     Accuracy: 85.39999999999999
```

vi. System Limitations

While sentiment analysis is useful, it is not a complete replacement for reading survey responses. In today's environment where we're suffering from data overload companies might have mountains of customer feedback collected. Yet for mere humans, it's still impossible to analyze it manually without any sort of error or bias.

It is impossible to analyze large amounts of data without an error.

vii. Conclusions

The system is tested for different reviews and the emotications are also successful for every review .It shows no of provided features, we are getting the accuracy of each algorithm we used . Different results show that the analysis gives good performance with ANN. The average accuracy of this system for test review is 85.399%

viii. Future Enhancement

Applications of sentiment analysis
will continue to grow in the future and that the implementation of sentiment analytical techniques will be standardized in various systems and services. The proposed future work will focus on three different characteristics chosen to investigate various datasets combining logistic regression and SVM algorithms. It can find unfair positive reviews and unfair negative reviews, reputation issues, and collusion and control through this work. The experimental method can study the accuracy, precision, and recall of both algorithms and can determine accurate and less time feature selection.

REFERENCES

[1] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, 2020, pp. 1-4, doi: 10.1109/ICCIS49240.2020.9257657.

[2] Wongkar, Meylan & Angdresey, Apriandy. (2019). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. 1-5. 10.1109/ICIC47613.2019.8985884.

[3] Wongkar, Meylan & Angdresey, Apriandy. (2019). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. 1-5. 10.1109/ICIC47613.2019.8985884.

[4] **Albert-based sentiment analysis of movie review**
Zhongxiang Ding;Yali Qi;Deping Lin
2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)

[5] "Sentiment Analysis of Review Datasets using Naive Bayes and K-NN Classifier"
Authors- Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose.
July 2019 DOI: 10.5815/ijieeb.2016.04.07

[6] "Sentiment Analysis of Review Datasets using Naive Bayes and K-NN Classifier"
Authors- Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose.
July 2019 DOI: 10.5815/ijieeb.2016.04.07

[7] IMDb Sentiment Analysis COMP 551 - Group 17
Authors : Beatrice Lopez ,Minh Anh Nguyen and Xavier Sumba
February 2019

[8] Kurniasari, Lilis & Setyanto, Arif. (2020). Sentiment Analysis using Recurrent Neural Network. Journal of Physics: Conference Series. 1471. 012018. 10.1088/1742-6596/1471/1/012018.