Text Classification for Sentiment Analysis Using Machine Learning and Deep Learning Algorithms, Case Study: Twitter Dataset

ABDULKHALEQ AYDIN A[1]

[1]*Department of Software Engineering, School of Engineering*
Ataturk University
[1]abdulkhaleqaydina.abdulkhaleq.aydin.a22@ogr.atauni.edu.tr

This Research Paper was Submitted in Fulfilment of The Requirements for Deep Learning and Convolutional Neural Networks Class
Prof.Dr. FERHAT BOZKURT

Abstract— **Sentiment analysis, also known as opinion mining, is a crucial task in natural language processing that aims to determine the sentiment expressed in textual data. In this paper, we present the results of a text classification task that was performed on the Sentiment140 dataset, a widely used dataset for sentiment analysis using Machine Learning and Deep Learning approaches. The algorithms that we used in this paper were 8 algorithms including "Logistic Regression, Random Forest, SVM, Decision Trees, Naive Bayes, KNN, ANN and MLP". Our approach is to use tweets polarity in order to do sentiment analysing, which can be used later for analysing any social media company or to view public satisfaction about a topic, subject, or even a company. This report provides an overview of the dataset, details the methodology used, and presents the results and findings of the sentiment analysis task also known as the "text classification tast".**

Keywords— **Machine Learning, Deep Learning, Sentiment Analysis, Text Classification**

I. Introduction

Sentiment analysis plays a crucial role in understanding people's opinions and attitudes towards various topics. With the advent of social media platforms, analyzing sentiment in the text has become increasingly important. Our objective is to highlight the potential of machine learning and deep learning algorithms to accurately classify these tweets into positive or negative sentiment categories and get the most insight of it whether to analyse customer satisfaction with a company or even to examine how the public waves about a topic or event in a certain time.

In this project, we delve into the fascinating field of sentiment analysis using the widely popular Twitter platform. Why Twitter Sentiment Analysis ? and what can be used for ? Well, it can be used for various applications like :

Market Analysing: Companies can use sentiment analysis to understand how people feel about their products or brands. And this can help them to improve their products and marketing strategies.

Brand Reputation Monitoring: Businesses can use sentiment analysis to monitor what people are saying about them online. This is crucial for helping them to identify and respond to negative feedback and fix the situation.

Political analysis: Sentiment analysis can be used to examine public opinion about political issues or candidates. In addition, to give helpful insights for political campaigns by understanding how people feel about different issues and targeting their messages to specific groups of voters.

Customer Service Analysis: Companies can use sentiment analysis to monitor customer feedback on social media and other online platforms. This info can be used by these companies to identify and respond to customer complaints and improve their customer service.

we focused on the Sentiment140 dataset, which was created by Alec Go, Richa Bhayani and Lei Huang, in addition to some books and papers about text classifications. The goal of our project is to develop a robust classification system that can effectively differentiate between positive and negative opinions expressed in text. By leveraging both machine learning and deep learning methods, we aim to accurately classify tweets based on the sentiment they convey.

Sentiment analysis plays a crucial role in understanding public opinion and sentiment towards various topics, products, or services. By automatically classifying tweets as positive or negative, we can extract valuable insights from the vast amount of user-generated content on Twitter. This has far-reaching implications for businesses, marketers, and researchers alike.

Our project utilizes a combination of traditional machine learning algorithms, such as Support Vector Machines (SVM) and Random Forest, Logistic Regression, Decision Trees, Naive Bayes, KNN , along with advanced deep learning techniques like Artificial Neural Networks (ANN) and Multilayer Perceptron (MLP). By employing these methods, we aim to achieve high accuracy and precision in sentiment classification, enabling us to uncover valuable patterns and trends within Twitter data.

Through this project, we aim to contribute to the field of sentiment analysis by showcasing the effectiveness of machine learning and deep learning techniques in classifying sentiment in Twitter data. By harnessing the power of these methods, we can gain valuable insights into public opinion, customer satisfaction, and market trends, paving the way for data-driven decision-making in various domains

II. The Sentiment140 dataset

The Sentiment140 dataset serves as the cornerstone of our project, offering a valuable collection of Twitter data specifically curated for sentiment analysis. This dataset was meticulously compiled by Alec Go and Richa Bhayani and consists of over 1.6 million tweets collected from the Twitter platform. Each tweet is associated with a sentiment label, indicating whether it expresses a positive or negative sentiment. In addition, this dataset offers a valuable resource for training and evaluating sentiment analysis models, enabling us to gain insights into sentiment patterns in social media data.

Comprising a vast number of tweets, Sentiment140 enables us to explore the opinions and emotions shared by users across a wide range of topics. Each tweet in the dataset has been annotated with sentiment labels, categorizing them as either positive or negative. This annotation process allows us to leverage supervised learning techniques to train our models and accurately classify the sentiment of unseen tweets.

One of the remarkable aspects of the Sentiment140 dataset is its relevance to real-world scenarios. Twitter, being a popular and sometimes influential platform, reflects the opinions, views, and emotions of millions of users worldwide. By analyzing this dataset, we gain valuable insights into public sentiment and the collective voice of individuals on various subjects.

III. Methodology

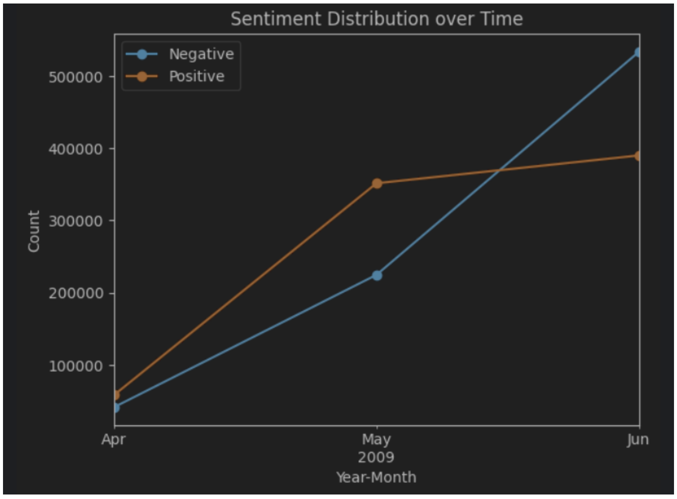
In our purse paper we can divide our methodology to two steps:

A. *Having a broader look at our data(EDA)*

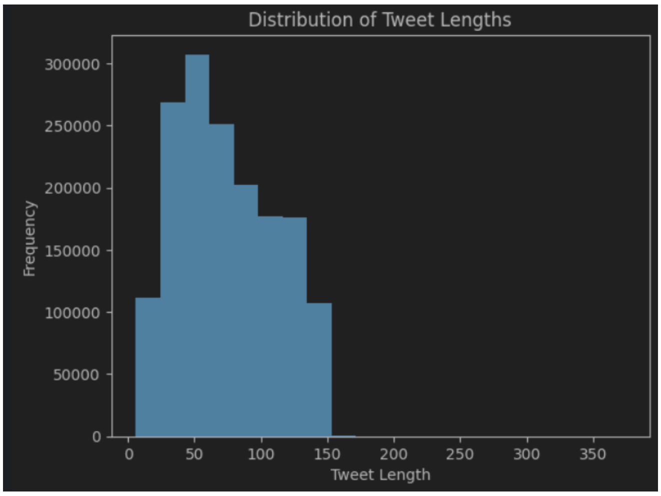
In any machine learning process, it should begin by analysing the data, what it contains, the features and behaviours toward each other, the variance, and correlations between the variables which affect the machine learning models.

In our Exploratory Data Analysis, we looked at the Distribution of Sentiment Classes which shows us each tweet's polarity per it counts. In addition, we examined some important sides like word count distribution which in our case shows us the length or size of the tweets present in the dataset.It's obvious that most of the tweets have a word count between 0 and 20. Also, Most Common Words: its important to know the most common words in the dataset as it will help us to understand the context of the tweets and also to perform feature engineering.

Average Word Length: this can help us to dive into the linguistic characteristics of our dataset, and understand the typical word length in the tweets and can provide information about the language usage and style. Sentiment Distribution over Time: this can provide valuable insights into the dynamics of public opinion on Twitter. It can help identify periods of heightened positive or negative sentiment, detect sentiment spikes or dips, and understand the sentiment patterns associated with different topics or events, and in our case study, we sight that there was a spike in negative sentiment in the month of April 2009 and a spike in positive sentiment in of June 2009. This could be due to some events that happened during that time period.

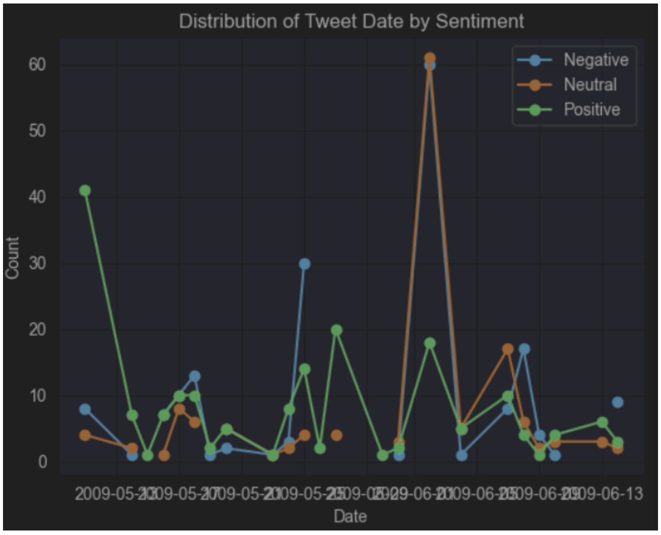


Distribution of Tweet Length, which consider one of the important features, which it can give us insight into user behaviour on the platform, for example, how the user are expressing their opinions and ideas within the tweet characters limit, another example is,imes the negative views tend to be longer in its length that the positives one.

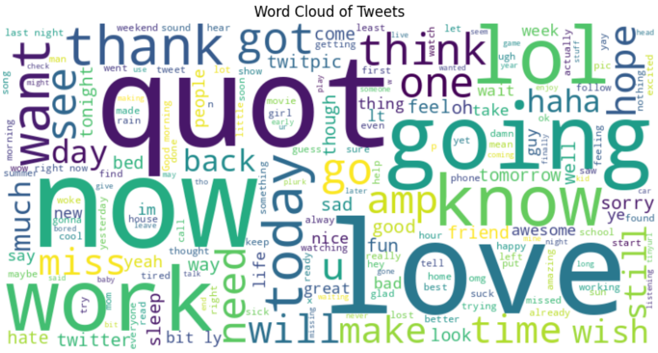


Distribution of Tweet Source which can give us a lens for identifying the platforms or applications from which the tweets originate. It provides information about the different sources through which users access and interact with Twitter. and understand the context and potential biases associated with the data.

Distribution of Tweet Date by Sentiment which consider one of the analysing metrics that help us to understand how sentiments fluctuate over time and whether there are any specific periods or events that impact sentiment. This metric can be useful for tracking public opinion, identifying sentiment shifts, or studying the impact of specific events on the public wave.



Distribution of Word Frequency using the word cloud library is critical to providing a representation for the most common words which can give us an overview of the language and vocabulary used by Twitter users in expressing their sentiments.In addition to give an insight into the understanding of words' importance, like giving the larger words indication of higher frequency which leads to easily identifying the most prominent and impactful words in the dataset.



B. Preprocessing and applying the proposed model's

Data Preprocessing: We conducted thorough preprocessing on the raw tweet data to remove noise and standardize the text. This involved cleaning the text by removing special characters, URLs, and mentions, as well as tokenizing and stemming the words to ensure a consistent representation.

Feature Extraction: To convert the preprocessed text data into a numerical format suitable for machine learning algorithms, we employed the TF-IDF (Term Frequency-Inverse Document Frequency) technique. This approach assigns weights to words based on their frequency and importance in the document.

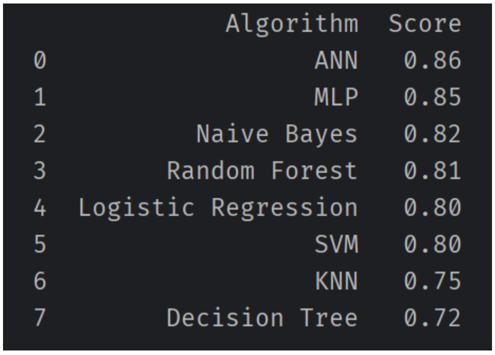
Model Selection: We experimented with several classification algorithms, including Logistic Regression, Random Forest, SVM, Decision Trees, Naive Bayes, KNN, ANN and MLP, to identify the most effective approach for sentiment analysis on the Sentiment140 dataset. Each algorithm was evaluated based on its performance metrics and suitability for the task.

Model Training and Evaluation: The selected classification algorithms were trained on the preprocessed dataset and evaluated using various metrics, including accuracy, precision, recall, and F1 score. These metrics provided insights into the models' ability to accurately classify tweets into positive or negative sentiments.

Hyperparameter Tuning: it was performed to optimize the performance of the selected models and accurate predictions.Grid search, a technique that explores different *combinations of hyperparameters, was employed to find the best configuration.*

IV. Results and Discussion

The sentiment analysis task on the Sentiment140 dataset yielded the following results:



```
             Algorithm  Score
0                  ANN   0.86
1                  MLP   0.85
2          Naive Bayes   0.82
3        Random Forest   0.81
4  Logistic Regression   0.80
5                  SVM   0.80
6                  KNN   0.75
7        Decision Tree   0.72
```

These results indicate that the Deep Learning approach performed well in classifying tweets into positive and negative sentiments, especially ANN . The high accuracy suggests that the model was able to make accurate predictions on the majority of the test dataset. The precision, recall, and F1 score further validate the model's performance, indicating a good balance between correctly identifying positive and negative tweets. In contrast, Decision Tree and KNN were the poorest performers among all the 8th algorithms. Finally, Hyperparameter tuning has played a crucial role in optimizing the model's performance and getting the optimum results from each model.

V. Conclusion

In conclusion, this sentiment analysis study on the Sentiment140 dataset demonstrated the effectiveness of various machine learning and deep learning algorithms in classifying tweets into positive and negative sentiments. Machine Learning algorithms like Logistic Regression, Random Forest, Decision Trees, Naive Bayes, KNN, SVM and Deep Learning ANN and MLP algorithms were evaluated, and their performance metrics were compared. Based on the results, ANN and MLP by far were the best among all the models, which emerged as the most promising approach for sentiment analysis on the given dataset, showing higher accuracy, precision, recall, and F1 score. The findings of this study contribute to the understanding of sentiment analysis techniques and their application in social media data.

VI. References

[1] To access the dataset : http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

[2] For source code about the project on my GitHub page: https://github.com/Dr-LazyMazy/Text-Classification-for-Sentiment-Analysis.git

[3] A. Go, R. Bhayani, and L. Huang, “Twitter Sentiment Classification using Distant Supervision.” Available: https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[4] Jacob, “Text Classification for Sentiment Analysis – Naive Bayes Classifier,” StreamHacker, May 19, 2010. https://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/

[5] C. M. Suneera and J. Prakash, "Performance Analysis of Machine Learning and Deep Learning Models for Text Classification," 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342208.

[6] Manning, C. D., Raghavan, P., & Schütze, H. (2009). An Introduction to Information Retrieval (Online edition ed.). Cambridge University Press Cambridge, England. http://www.informationretrieval.