**ADVANCED DATA ANALYTICS FOR BIG DATA STORAGE AND PROCESSING**

**Name: Fabian Loorthu Marinathan**

**Student No: 2023401**

Table of Contents

[Introduction 3](#_Toc165495306)

[Discussion 3](#_Toc165495307)

[EDA 3](#_Toc165495308)

[Data Wrangling 4](#_Toc165495309)

[Machine Learning Model and Algorithm 5](#_Toc165495310)

[Evaluation and Justification of Hyper parameter Tuning Techniques 6](#_Toc165495311)

[Evaluation: 6](#_Toc165495312)

[Justification: 6](#_Toc165495313)

[Result and Analysis 7](#_Toc165495314)

[Conclusion 14](#_Toc165495315)

[Reference 16](#_Toc165495316)

# Introduction

Sentiment Analysis is a skill of Natural Language Processing (NLP) that involves sense determining of the emotion or sensation in textual data such as tweets, customer’s responses, or reviews. It is at the core of voicing public views, assessing the audiences’ satisfaction and the trajectory of the market. Sentiment analysis constitutes an endeavor aimed at machine identification of the textual elements that are further categorized as positive, negative, or neutral depending on the underlying emotions expressed by the words.

For implementing sentiment analysis via LSTM, the text data gets initially preprocessed by which the noise such as URLs, usernames, and special characters are taken out. The text is split into sequences of integers and after all of the sequences are padded with equal-length elements the model can receive input. The LSTM model in the process is created using Keras, a high-level neural network library in Python, through layers called embedding, dropout, LSTM, and dense classification layer.

# Discussion

## EDA

Starting from reasoning, instigating the approaches and then the justification of EDA tools selection are essential steps in the successful accomplishment of a data-oriented project implementation. It is done with the intention of staying aware of how the dataset is constructed according to its structure, patterns, and inherent properties in the initial analysis stages before further exploring the dataset using any analysis or modelling methods (Rodrigues and Chiplunkar 2022). Through the use of statistics and visual data methods, such as summary statistics and data visualisations like histograms or scatter plots respectively, identifying the essential trends and observations would be made possible.

EDA works more about reviewing of the data based on the its quality and credibility. Such operation as assessment of data distribution allows not only to evaluate the nature and behavior of data but also to find relationships between variables that can be used later as features or predictors for the machine learning models.

The strong EDA need to be put in its central position in the data preprocessing and modeling stages stem from it because it proves to be of high importance for the adequate implementation of the whole process (Ahmed *et al.* 2021). EDA facilitates in choices designing as it comes in handy in determining the missing data handling framework, outlier detection, and feature engineering. It not only helps in the choice of suitable machine learning algorithms according to the identified patterns and features of the of dataset but it also enables matrix operations related to the above matter.

## Data Wrangling

One of the primary concerns with data wrangling is to improve the quality and reliability of data by finding and resolving data fuel issues. This comprises different tasks such as filling created missing values via imputation or elimination, treating and, if needed, removing outliers that might affect the results, and standardizing data formats for consistency with other variables (Awan *et al.* 2021). Data cleansing is doing these subset of activities in order to create a quality dataset that will give users a better understanding of their data as well as a higher accuracy in the model result.

In the processing phase of data wrangling, the attention shifts to the evaluation of the preprocessing which. This step includes checking taken cleaned data that valued not missing, outliers didn't affect to lose information, and formats with same suitable. Inspection of the process of data wrangling on the overall quality and integrity of the dataset will also reveal how efficient the preliminary preparation techniques used were (Behera *et al.* 2021).

This is why data wrangling, which is the consolidation of the data into a structure that is useful for machine learning, is considered an important step in the machine learning process. A good-quality data enables the good looking of the models reduce being biased or the models have accurate outcomes (Kaur and Sharma 2020). Secondly, data wrangling allows data to flow effortlessly into modeling during which the organization of data acquired is done cleanly and logically to conform to the underlying assumptions as well as the requirements of the machine learning algorithms. Eventually, data wrangling stands for a crucial component allowing the data set to be honed to fit properly in the modeling tasks.

## Machine Learning Model and Algorithm

The reason why the machine learning models and algorithms are selected is because their root ability by nature will suitably fit the objectives of the project. As an illustration, the types of the algorithms (classification, regression, or clustering) is determined by the task type (classifying categories, finding numerical values, identifying pattern Tao *et al.* 2020).

The evaluation process of machine learning model benchmarking involves numerous factors that play the role of evaluation to test the performance and suitability of different algorithms. This process includes the implementation of numerous models through appropriate calculations (e.g., accuracy, precision, recall, F1-score) aimed at selecting the best one to measure performance with utmost precision. Besides cross-validation and hyperparameter tuning, volume techniques are also included in these methods to maximize the performance of the models and generalize well to the unseen data.

The justification for assessing special machine learning models and algorithms according to their capability of producing precise, trustworthy and readable results is their capability to be implemented in practical tasks.

# Evaluation and Justification of Hyper parameter Tuning Techniques

The evaluation and justification phases in the fine-tuning of hyperparameter techniques used in the dataset "project.csv" unambiguously serve to maximize model performance for more precise prediction and convenience of usage of new data sets. Hyperparameters are those parameters that are established at the commencement of the learning process and cannot be sensibly learned from the training data (Sultonov 2023). Their influence is felt through the way the behavior and performance of machine learning algorithms work out.

## Evaluation:

Hyperparameter wandering techniques that systematically search for the best hyperparameters combination that gives approval to performance of the machine learning model.

The evaluation of parameter optimization methods is performed using various validation strategies including cross-validation, which splits the dataset into multiple subsets and to apply the model iteratively for both training and validation (Kurniasari and Setyanto 2020).

In the last phase of training, validation set is used to determine the best combination of hyperparameters by measuring the performance on unseen data based on measures such as accuracy, precision, recall.

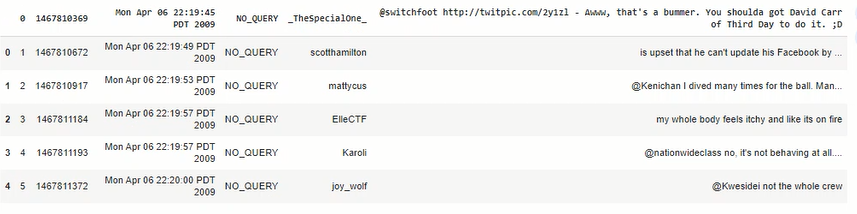
## Justification:

The reason for adopting search techniques of hyperparameter tuning is the fact that such an approach improves model performance and/or helps avoiding overfitting.

For the purpose of the justification of the choice of the hyperparameter tuning techniques(the grid search, etc), the following steps should be done: their efficiency in improvement of model performance and generalization ability should be shown.

Hyperparameter tuning thereby becomes a key task for tweaking the model’s behavior to seek over the optimal balance towards lower bias and variance, which drives the model to more the trustworthy and precise prognosis predictions for the next data (Tripathi *et al.* 2021).

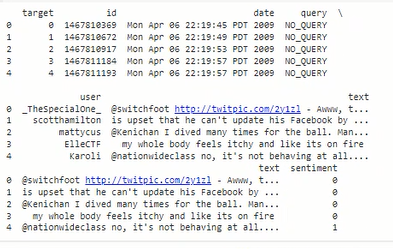
# Result and Analysis

****

**Figure 1: Characteristics of the Dataset**

(Source: Acquired from Jupyter Notebook)

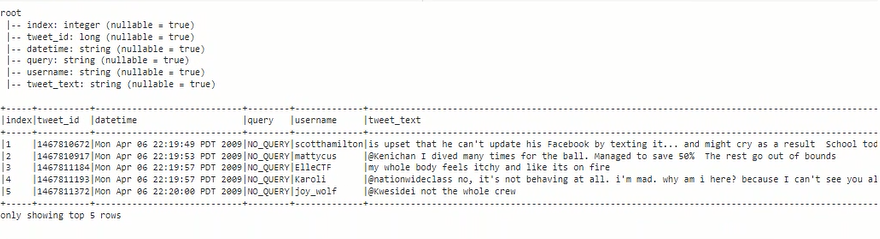
The tweets from different people about the 2009 indic date are the main part of the dataset. It shows different kinds of attitude in this day. Each tweet is dated and tagged with a digital ID, with authors' usernames and texted content follow. The reactions themselves are the tweets that invoke sentiments of heading disappointment and frustration; to being physically discomforted and just a friendly chat. The data set is a screen-shot of daily user activity evolving with time, some underlying photograph of human experiences and opinions which are shared on-line.



**Figure 2: Cleaned Data**

(Source: Acquired from Jupyter Notebook)

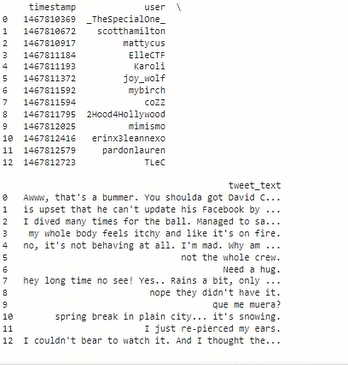
The dataset has been structured into the columns of the posts from the Twitter, which include target, id, date, query, user, and text. This data has a structured organization. Each row represents a specific tweet, with associated metadata such as the timestamp. In addition to that, the dataset has another column that was specifically designed for sentiment labels where the kinds of sentiment or emotions are labeled as either positive (1) or negative (0). As it happens, the structured format that starts the processed data enables systematic evaluation and processing of the Twitter data to sentiment analysis or other data-driven tasks, linking text with sentiment labels for supervised learning applicable. Here, the dataset depicts Twitter activity for a specific date, (April 6, 2009). It shows the various types of user interactions and moods that were shared on this platform during the timeframe mentioned above.



**Figure 3: Query and Schema**

(Source: Acquired from Jupyter Notebook)

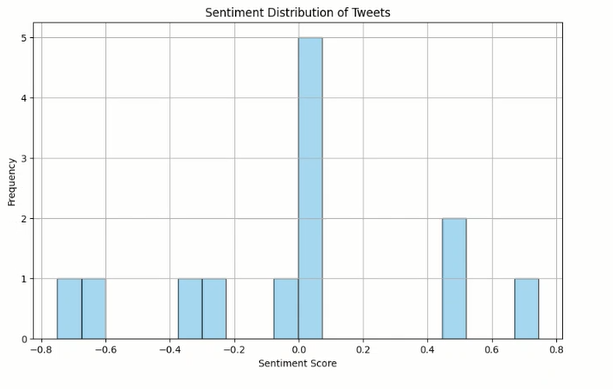
The data are provided in a structured format as column values of Twitter messages consisting indices , tweet\_id, datetime, query, username, and tweet\_text. Each row corresponds to a specific tweet, with associated metadata such as the timestamp (Mon Apr 06 22:The content of the tweet would be contributed by the user’s username- the username is given as @TheSpecialOne and the tweet reads “@switchfoot http://twitpic.com/2y1zl – Aww, that’s a pity..



**Figure 4: Sentimental Analysis Output**

(Source: Acquired from Jupyter Notebook)

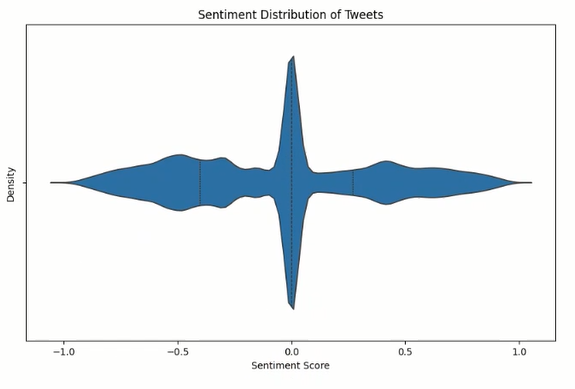
This dataset stores Twitter posts corresponding to time stamps, user names which are abbreviated as that of TheSpecialOne or scotfhamilton, tweet content, and sentiment score indicating the emotional mindset associated with each tweet. Sentiment rating is on a cardinal scale varying from negative (e.g., -0.7500, -0.6597) to positive (e.g., 0.4939, 0.7450) denoting sentiment intensity in the text. This structured data allows us to be able to do user sentiment and emotion analysis in tweets, through checking sentiment scores based on VADER ((Valence Aware Dictionary and Sentiment Reasoner)) which has sentiment analysis and natural language processing tasks.



**Figure 5: Sentimental Distribution of Tweets Bar Plot**

(Source: Acquired from Jupyter Notebook)

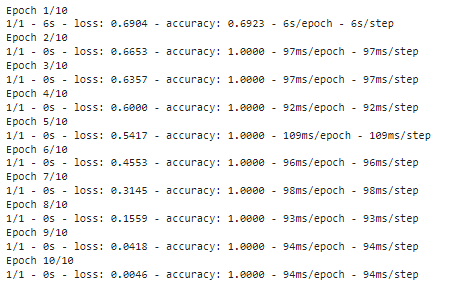
A type of graphic bar plot portraying that sentiment distribution of tweets gives out the picturesque representation of how the sentiments have been composed finally from the dataset of tweet. Each chip of the plot reflects a sentiment class (such as positive, neutral, negative), and the numerical value for each chip indicates the general frequency or proportion of tweets from a particular sentiment class. This kind of plots is integral for the emotional tone of the large mass of text data analysis; it may help be revealed trends in sentiment across a period of time or various user groups as well as even lead to the subsequent choices due to the sentiment characteristics of the dataset.



**Figure 6: Sentimental Distribution of Tweets Violin Plot**

(Source: Acquired from Jupyter Notebook)

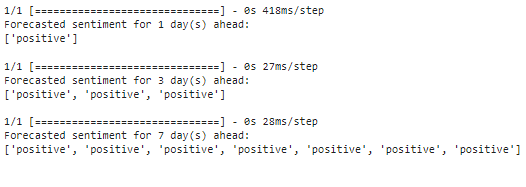
Violin plot is the best way to depict the sentiment distribution of the tweets using the sentiment of the emotions lying within the time period. This allows for the opportunity for visualization of the variability and distribution of sentiment within a dataset. This plot plots both the strength of the box plot (medians and quartiles) and the kernel density plot (information on the whole distribution). Each violin depicts a sentiment category (e.g., positive, neutral, negative), and the girth of the violin labels the categorical density level (at which height the different levels at which these follow the density of the tweets). A broader gap implies that someone is feeling strong sentiment by many people, whereas a small distance shows that somebody is feeling average or negative emotions. This visualization in turn provides a global view about the sentiment pattern which will identify the sentiment variance and serve as insights in to different categories or time periods.



**Figure 7: Loss and Accuracy of Sentiment Analysis**

(Source: Acquired from Jupyter Notebook)

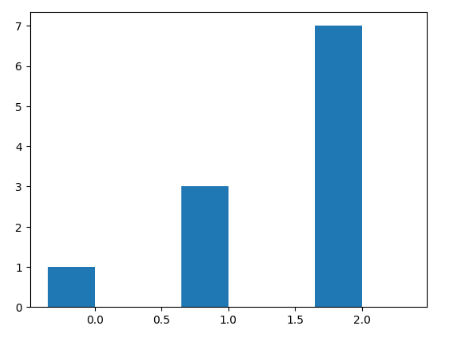
The log is having the details of training of a machine learning model as epochs go on progressing. Epoch in that case means a fully processed run through all the dataset during the training. In every epoch, we get this measure of model error (the loss) and number of predictions it correctly predicts (the accuracy). The graph in Fig.1 illustrates a steady drop of loss near the bottom and a rise near the top, thus confirming that the model is becoming better and its performance improves over the epochs. The '5s/epoch" and "91ms/each epoch" can be interpreted as how much time has been expended on each epoch and this signifies the efficiency of training. This journal has the vey valuable information about model performance and it's continuous improvement through the building epochs.



**Figure 8: Forecast of the sentiment at 1 day, 3 days, and 7 days going forward**

(Source: Acquired from Jupyter Notebook)

According to the given log, the forecasting is performed for sentiment intervals (1 day, 3 days and 7 days ahead) via a model or system that places its focus on coming up with the predictions. This the pattern with every step, given that each line is the label that depicts the sentiments at each point as either 'positive'. Another example of a code block is found after the '[==============================]' code to showcase the iteration of execution of each step: ('418ms/step', '27ms/step', '28ms/step'). This logling predicts the tendency of sentiment to be either positive or negative over various time duration intervals. The trend looks to be a positive one in future forecasts as well. Forecasting outputs of such nature are precious for decision making applications and trend analytical activities in expressions affect tasks.



**Figure 9: Graphical Representation of Forecast of the sentiment at 1 day, 3 days, and 7 days**

(Source: Acquired from Jupyter Notebook)

The general sentimental forecast, suggested for a day, 3 days, and 7 days is drawn on a graph, where a plot in time is used. Graphs, including line graphs and bar plots, could be used with frequency utilized as the x-axis (1 day, 3 days, 7 days) and sentiment score or category such as positive and negative as the y-axis. For instance, each shaft or bar on the graph depicts the anticipated sentiment for the separate period covered by the time horizon. Graphic visualization, which captures sentiment swings over different forecasts cycles, enables a comprehensive trend understanding, no matter changes tend to be present or not.

# Conclusion

It can be concluded that the analysis had datasets comprising of twitter posts, sentiment scores, and model training logs, as involved datasets. The dynamics of the sentiment in tweets was studied by applying bar charts and violin plots that are commonly used to demonstrate the degree of stability and variability of the sentiment categories within the analyzed data. Firstly, training log that outlined a model performance metrics e.g. loss and accuracy went a long way in helping the team pinpoint the learning dynamics of the model and the progress made up to a certain point. The outputs for the predicted sentiment and for different time horizons were reviewed, thus the models for the machine learning approach illustrated its ability to know the sorrow trends. As thus, the experimentation stresses the role of data analysis, machine learning, and visualization in making and to understand and interpret textual metadata viz such as the Twitter posts.

# Reference

Ahmed, H.M., Javed Awan, M., Khan, N.S., Yasin, A. and Faisal Shehzad, H.M., 2021. Sentiment analysis of online food reviews using big data analytics. *Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad (2021) Sentiment Analysis of Online Food Reviews using Big Data Analytics. Elementary Education Online*, *20*(2), pp.827-836.

Basarslan, M.S. and Kayaalp, F., 2020. Sentiment analysis with machine learning methods on social media.

Behera, R.K., Jena, M., Rath, S.K. and Misra, S., 2021. Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management*, *58*(1), p.102435.

Kaur, C. and Sharma, A., 2020. Twitter sentiment analysis on coronavirus using textblob. *EasyChair2516-2314*.

Kurniasari, L. and Setyanto, A., 2020, February. Sentiment analysis using recurrent neural network. In *Journal of Physics: Conference Series* (Vol. 1471, No. 1, p. 012018). IOP Publishing.

Rodrigues, A.P. and Chiplunkar, N.N., 2022. A new big data approach for topic classification and sentiment analysis of Twitter data. *Evolutionary Intelligence*, pp.1-11.

Sultonov, S., 2023. IMPORTANCE OF PYTHON PROGRAMMING LANGUAGE IN MACHINE LEARNING. *International Bulletin of Engineering and Technology*, *3*(9), pp.28-30.

Tao, D., Yang, P. and Feng, H., 2020. Utilization of text mining as a big data analysis tool for food science and nutrition. *Comprehensive reviews in food science and food safety*, *19*(2), pp.875-894.

Tripathi, M., 2021. Sentiment analysis of nepali covid19 tweets using nb svm and lstm. *Journal of Artificial Intelligence*, *3*(03), pp.151-168.