Examining Brand Perception Challenges: VADER Approach vs. Traditional Analysis for Improved Understanding

V Karthick1, Vidhiya S B2, Vaisharli S3

*123Computer Science and Engineering Department,*

*Rajalakshmi Engineering College Chennai*

*1karthick.v@rajalakshmi.edu.in*

*2210701306@rajalakshmi.edu.in*

*3210701296@rajalakshmi.edu.in*

***ABSTRACT -* In today’s digital era , Social media plays a key role in Brand’s reputation and perception. Using a Data driven approach, It is easier to analyze the brand’s public review and where it stands among all the organizations. It helps the brand to understand and analyze their position and helps them improve their brand. Sentiment Analysis is one of the NLP** [[1]](https://paperpile.com/c/ISzt1k/UUkK) **techniques used to analyze positivity, Negativity or neutrality of a data, which helps in accurate brand perception. It is impossible for an individual to check lakhs and lakhs of comments on social media** [[2]](https://paperpile.com/c/ISzt1k/Ovq8) **and it is difficult to analyze the reputation of the brand among the public. This paper focuses on VADER (Valence Aware DIctionary and Sentiment Reasoner). classification which is a tool used for Sentiment Analysis within NLTK. VADER is a lexicon oriented sentiment analysis tool which can be used to analyze the customer reviews which are all over on the social media which can be uploaded as a csv to the User Interface and the user can understand how customers perceive a particular brand. The traditional approaches include lexicon based or model based or rule based. But Vader is a combination of Sentiment Lexicon and Rule Based model. VADER is much preferred for social media sentiment analysis since it is very fast and much more efficient than traditional methods. VADER has a f1 score of 0.96 and the lexicon based approaches has an**

**F1 score of 0.85. The rule based methods can achieve a f1 score of 0.9 when the rules are clearly defined which cannot be precise for social media texts. The objective is to analyze and know how a brand is grasped among the people. This approach harnesses the power of VADER which is a part of NLTK ,which is used to scrutinize and analyze social media conversations.Naive Bayes Classifier has an accuracy of 80-90 depending on the type of text.These help the business in targeted marketing , personalized brand improvement.**

**The Users can upload their data through the Streamlit** [[3]](https://paperpile.com/c/ISzt1k/AQyJ) **UI. The data is loaded as dataframes using pandas. The text in the data frames are transformed into strings. VADER from NLTK is applied to each text and it gives a sentiment score which ranges from -1(negative) to 1(positive). Then they are classified into Positive , Negative and Neutral. The reviews are combined to know the overall sentiment distribution (positive , negative or neutral) and the pie chart shows the sentiment distribution of a brand. The line chart combines the text data according to the month they were created. The text data without stop words are visualized using word clouds.**

***Keywords - Brand Perception, Sentiment Analysis, VADER, NLTKz***

**1. INTRODUCTION**

Brand perception analysis plays a major role in analyzing the public’s review of a product or service of an organization. Nowadays, Social media plays a major role in collecting the data and from that data it is easier to analyze the product’s worth. Data is collected from social media like twitter. Short messages in twitter are called tweets which are in length of 148 characters and it is easier to analyze. Using this approach, as a commoner we can know the product's worth and the quality of that brand. It helps the e-commerce platform to understand the sentiment of the customer towards the product and it can help to improve the recommendation of a product among the customer and it is used to optimize the overall recommending strategy for a customer. This approach helps the organization to analyze their position in the market by understanding the customer satisfaction of that product. We can use this approach in the area of entertainment like media companies or movie studios even in music labels to know the audience sentiment about the content they create and also analyze the popularity of that media.

This approach includes the 1) VADER[[4]](https://paperpile.com/c/ISzt1k/V3Uh) classification technique which is a tool used to analyze the text sentiment in social media like twitter, instagram. 2) NLTK[[5]](https://paperpile.com/c/ISzt1k/jeLd) is a library of NLP that provides tools and algorithms for NLP tasks like tokenization, stemming etc. The basic idea of this approach is to analyze the public’s opinion about the product and categorize them into “positive”, “negative” or “neutral”.

Ankita Gandhi , Kinjal Adhvaryu , Soujanya Poria , Erik Cambria and Amir Hussain work on a project called Multimodal sentiment analysis[[6]](https://paperpile.com/c/ISzt1k/GBLH). ‘Multimodal’ refers to the multiple modes of communication. It also includes the various forms of medium like audio, video, and in the form of text to enhance interaction experience. Sentiment analysis grows much in the field of artificial intelligence for analyzing sentiment about a product or services of an organization. Opinions and ideas are growing towards sentiment analysis. For this purpose, multimodal sentiment analysis focuses on collecting data in the form of video instead of text alone.

Cheng Chen, Bin Xu , Jong-Hoon Yang, and Mi Liu did a project called “Sentiment analysis of Animated Film reviews using Intelligent Machine Learning”[[7]](https://paperpile.com/c/ISzt1k/NbPO). This project aims to analyze the sentiment of users towards the animated movie, the dataset which is used to train the model is collected from various users by analyzing their behavior. Even in a social media or e-commerce platform uses the information of the users browsing history, customers favorites, add-ons, most viewed content and liked content based on this the recommendation of a content for a user is done. This paper focused on the construction of several deep learning models [[8]](https://paperpile.com/c/ISzt1k/Y4CM), intelligent machine learning-based text sentiment classification. These models were compared based on their experimental results and methods. This analysis towards the movie only achieves the overall sentiment of the review data and fails to analyze the specific context in the movie, specifically plot, special effects and content of the movie.

**3. MATERIALS AND METHODS**

The dataset used to train the model is in the form of CSV(Comma separated values) [[9]](https://www.researchgate.net/publication/344969402_Pytheas_pattern-based_table_discovery_in_CSV_files) files, even users can also use their own dataset to train the model for the particular dataset of a product to analyze the sentiment towards that product. The user also uploads their dataset which is in the form of a CSV file to train the model with different dataset of a particular product. This model uses VADER classification technique to determine the sentiment expressed in a text and also tells about how positive and negative it is. NLTK is a python library that provides algorithms for NLP tasks.

Hardware Requirements

Operating System(Windows)

RAM(4 GB )

Secondary Storage (256 GB Minimum)

Software Requirements

Internet Browser(Chrome/Edge)

Stable Internet connection

**4. EXISTING ALGORITHM**

**2.1 Support Vector Machines**

Using a Support Vector Machine for the analysis of sentiments [[10]](https://paperpile.com/c/ISzt1k/n04i) is an effective technique in its ability to discern non linear boundaries. It is effective since it is good in binary classification by finding out the optimal hyperplane which separates different classes. It processes textual data by converting them into numerical features through a bag of words or TF-IDF [[11]](https://paperpile.com/c/ISzt1k/ahlb). They describe the presence or absence of particular words or phrases in the text. It involves choosing a kernel function which includes linear , polynomial and radial basis function kernels.

**2.2 Long Short Term Memory**

The LSTM networks [[12]](https://paperpile.com/c/ISzt1k/cg0d)are variants of RNNs which particularly address the vanishing gradient issue , which happens in long sequences of data. It incorporates memory cells to selectively retain and forget . It allows to effectively get log- range dependencies in sequential data.

It consists of many memory cells, each has an input gate, a forget gate and an output gate. It can understand the context of a word with regards to its nearest words. It is therefore necessary to understand and analyze a piece of text. They have the capability to automatically extract the features.It is particularly useful in applications where the temporal aspect understanding is very crucial.

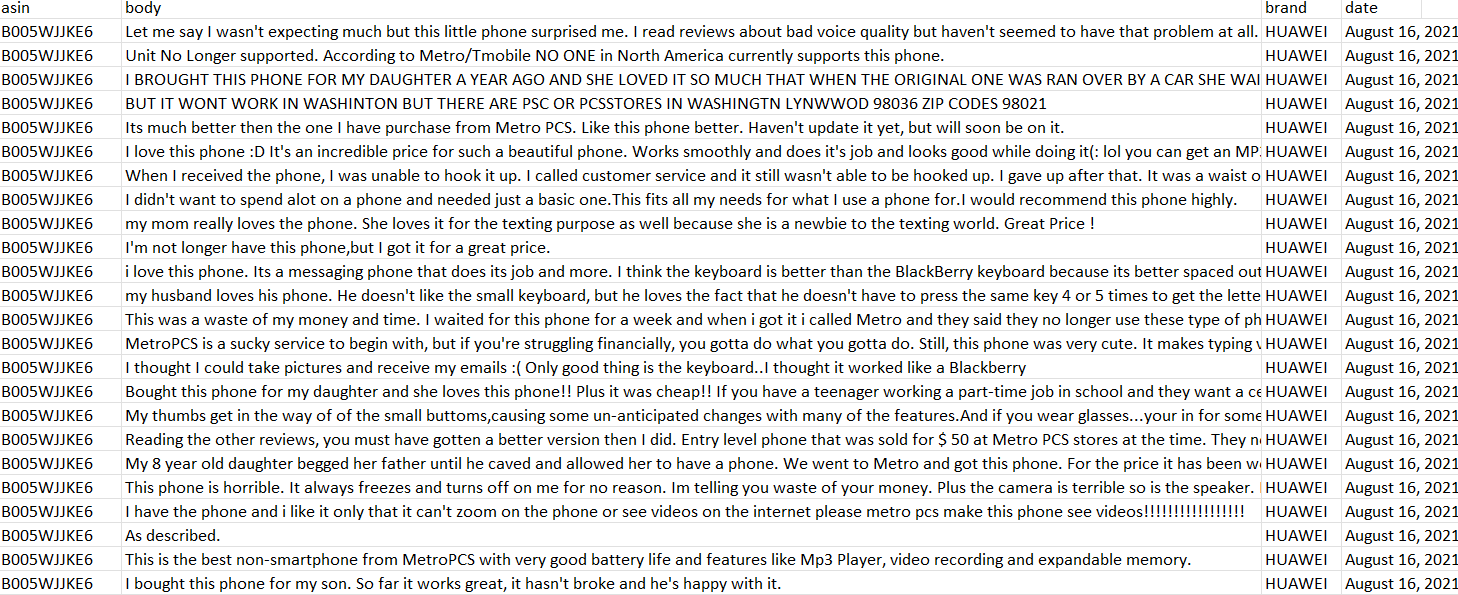
**2.3 Naive Bayes Classifier**

This technique is more efficient and has low memory requirements. It is also straightforward to implement and understand. It is well scalable to large datasets. It usually shows high bias but low variance which makes it less vulnerable to overfitting. It is more advantageous when the objective is to generalize to unseen data. It can be applied to both binary(positive/negative) and multiclass(positive/negative/neutral) sentiment analysis tasks [[13]](https://paperpile.com/c/ISzt1k/lJF5).

**4 PROPOSED SYSTEM**

**4.1Dataset**

The Mobile reviews dataset is publicly available on kaggle repository which contains asin , body ,brand and date columns and about 1063 records.



**4.2 Valence Aware Dictionary for Sentiment Reasoning**

Libraries like Streamlit(for UI) , Pandas , NumPy, Plotly Express [[14]](https://paperpile.com/c/ISzt1k/fZSr), Wordcloud , Matplotlib and NLTK(for VADER) are implemented in this approach.

It is specifically designed to analyze the sentiment especially in social media which often has slang words , emojis and misspellings.

Valence Aware Dictionary for Sentiment Reasoning analyzes by assigning a polarity score to each part of the text. It denotes the positivity , negativity or neutrality of the sentiment in the text.

**4.3 Normalization**

where α = Normalization constant (default value: 15) and x = sum of valence scores of constituent words.

The compound score is calculated by adding up all of the lexicon's word valence ratings, adjusting for rules compliance, and then normalizing the result to fall between the most extreme positive (+1) and negative (-1). If you're looking for a single, unidimensional sentiment measure for a particular statement, this is the most helpful metric.

**4.3 Model Architecture**

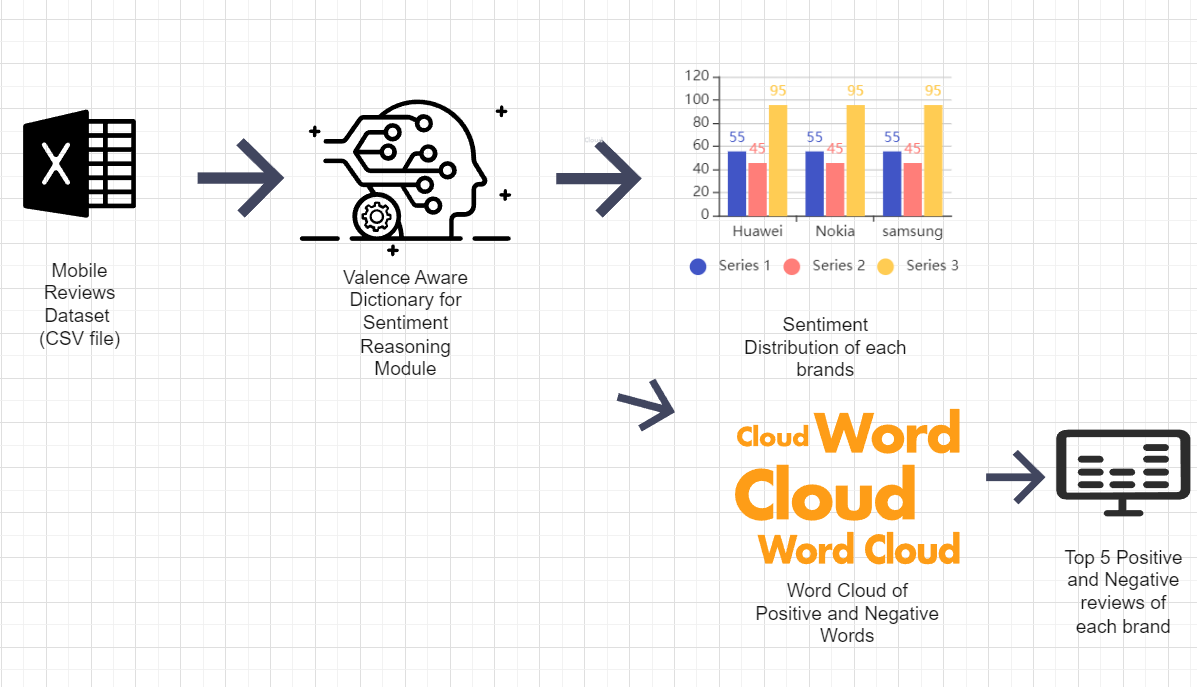


Fig 4.3.1 Model Architecture

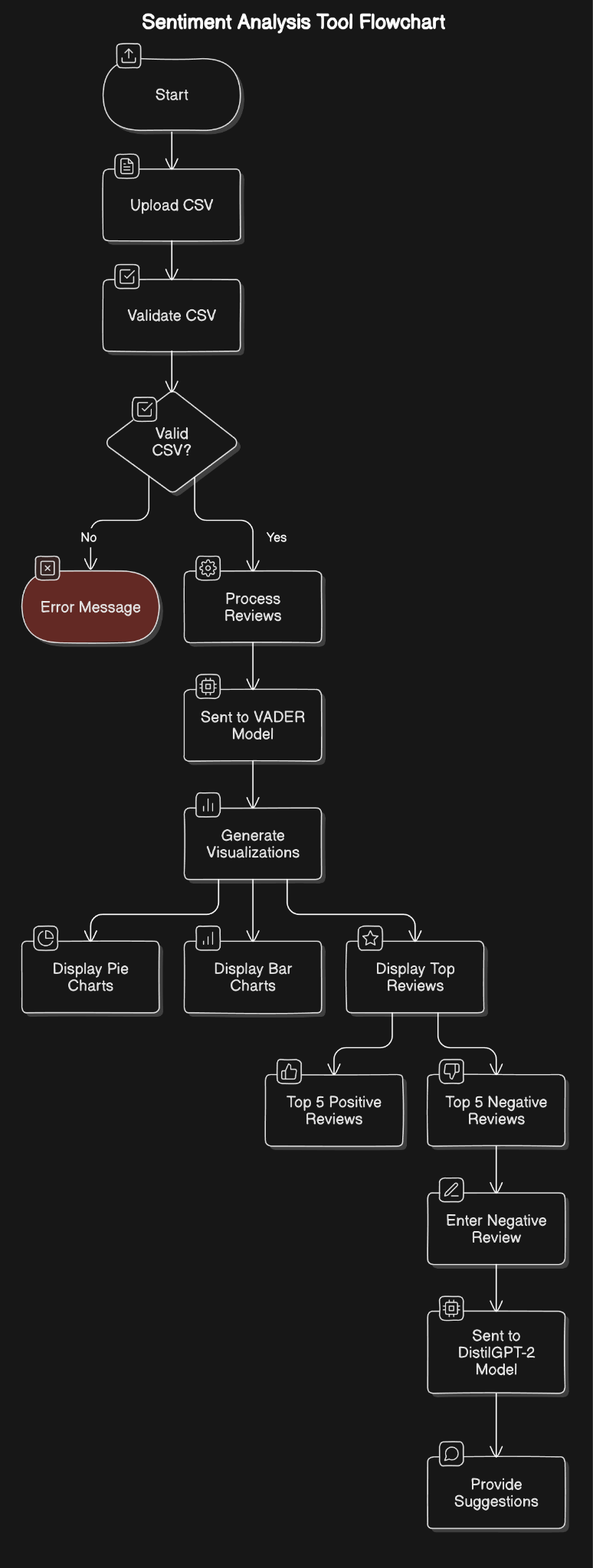
The dataset is uploaded into the UI module. The data is then transformed into Pandas DataFrame.Then the preprocessing is performed where all the letters in the text are converted into lowercase .

The Model then assigns a polarity score for each review and generates visualizations.

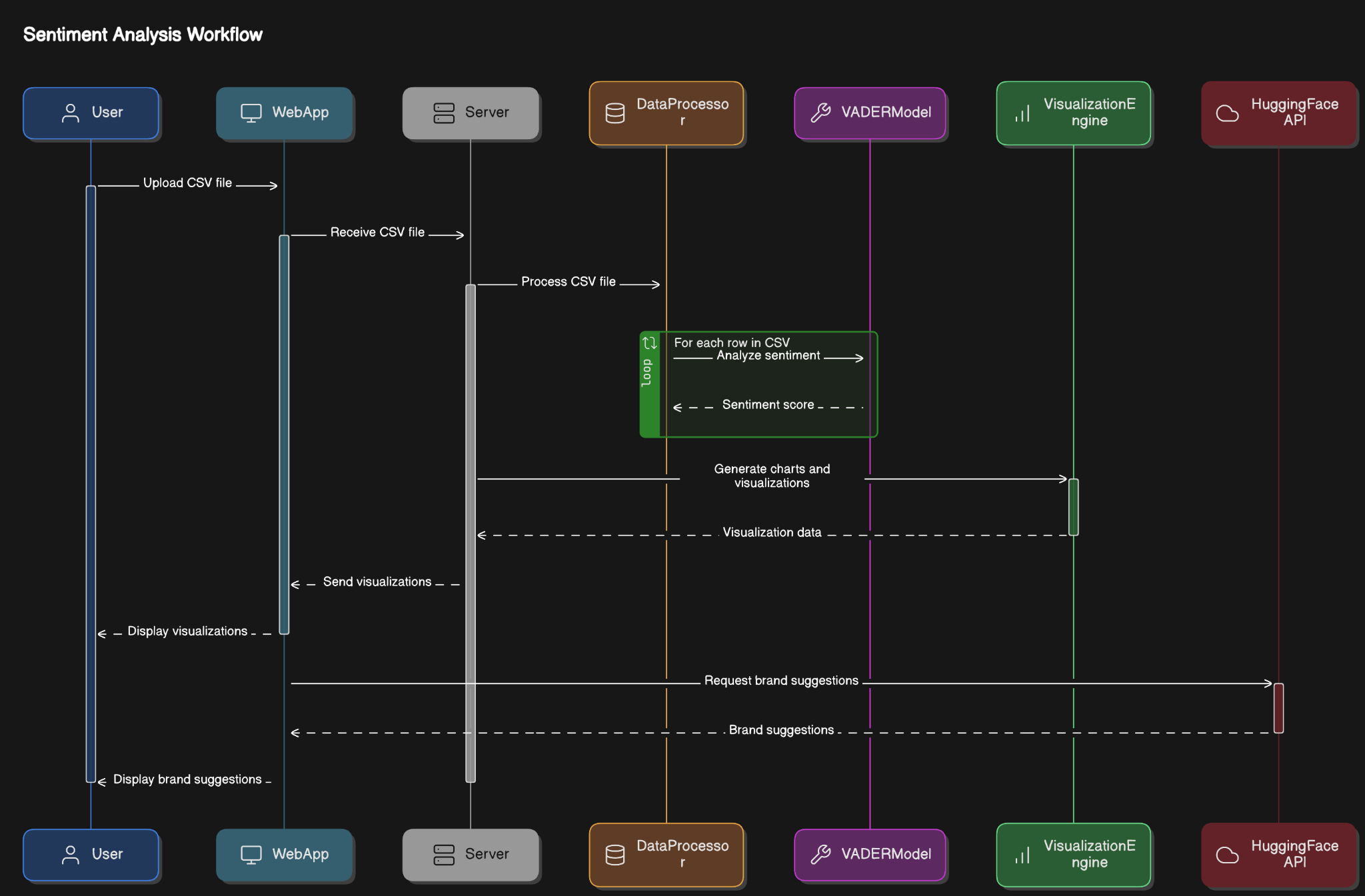
These Visualizations include Distribution by count and word clouds for positive and negative reviews.

The top 5 positive and negative reviews are displayed based on the polarity score.

**Proposed System**

****

**Fig 4.3.1 Process Flow Diagram**

****

**Fig 4.3.2 Sequence Diagram**

Then the sentiment analysis is performed through VADER SentimentIntensityAnalyzer to obtain a compound sentiment score and sentiment labels(positive , negative or neutral) are assigned according to the score.

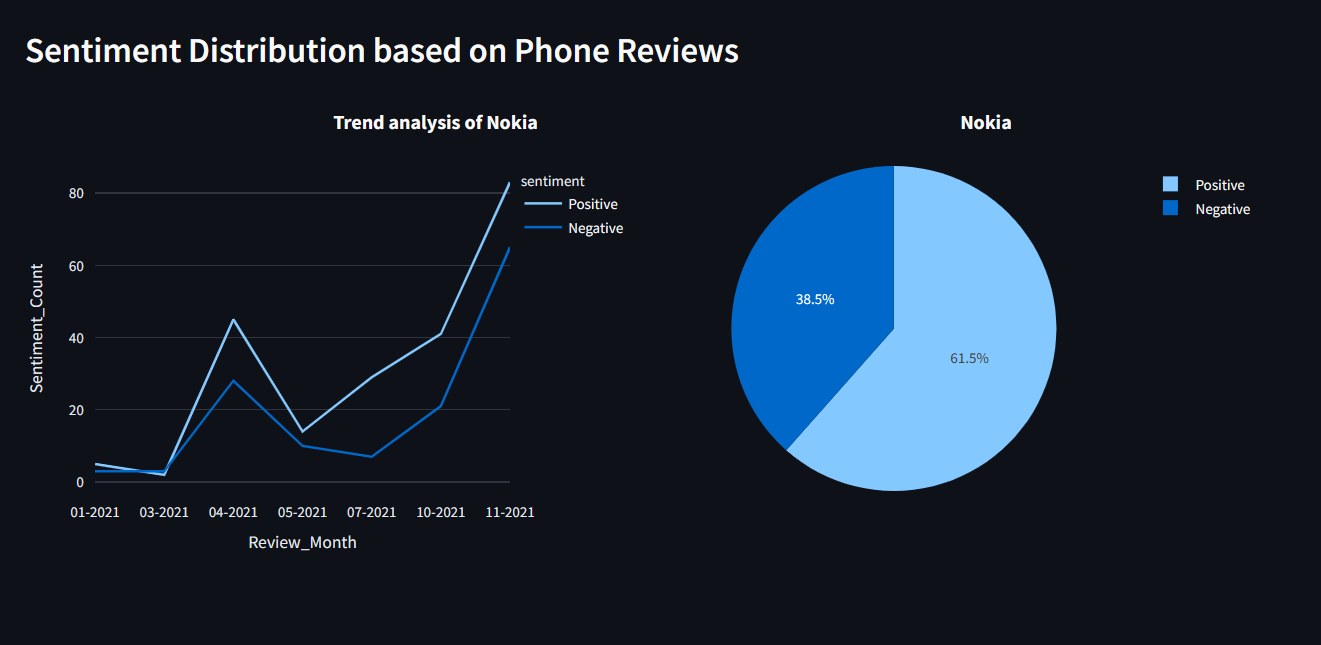
Then the data is aggregated to generate visual insights of sentiment distribution and trends using histograms and pie charts.

Word clouds are generated to visualize the most frequent positive and negative words where the stopwords are excluded.

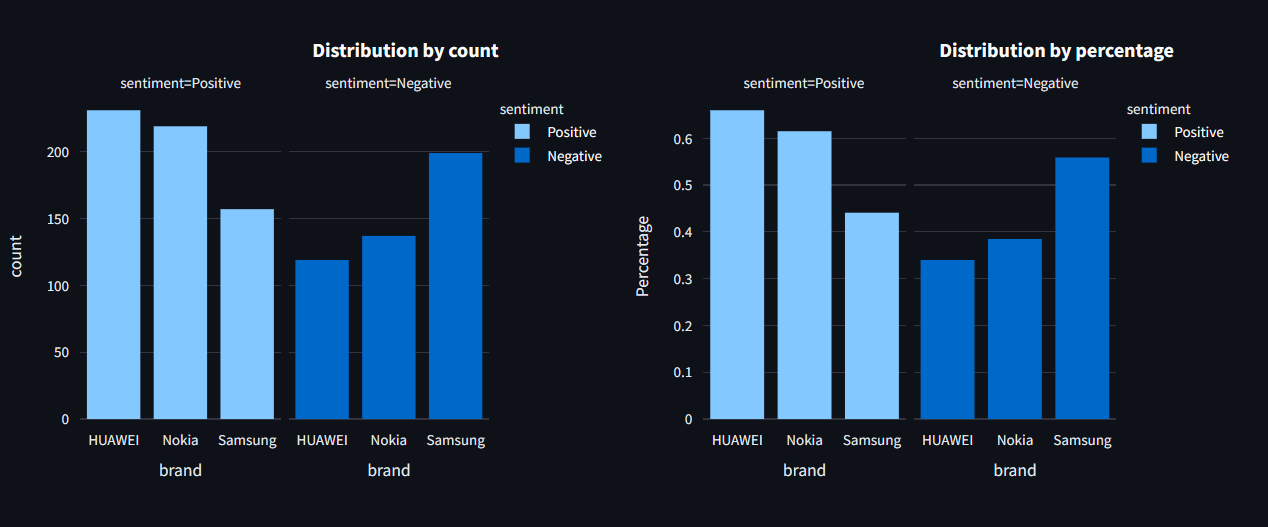
The top 5 positive and negative reviews are displayed on the Streamlit User Interface.

The suggestions are also included by using Hugging face transformer distil-gpt2.

**5 RESULTS**



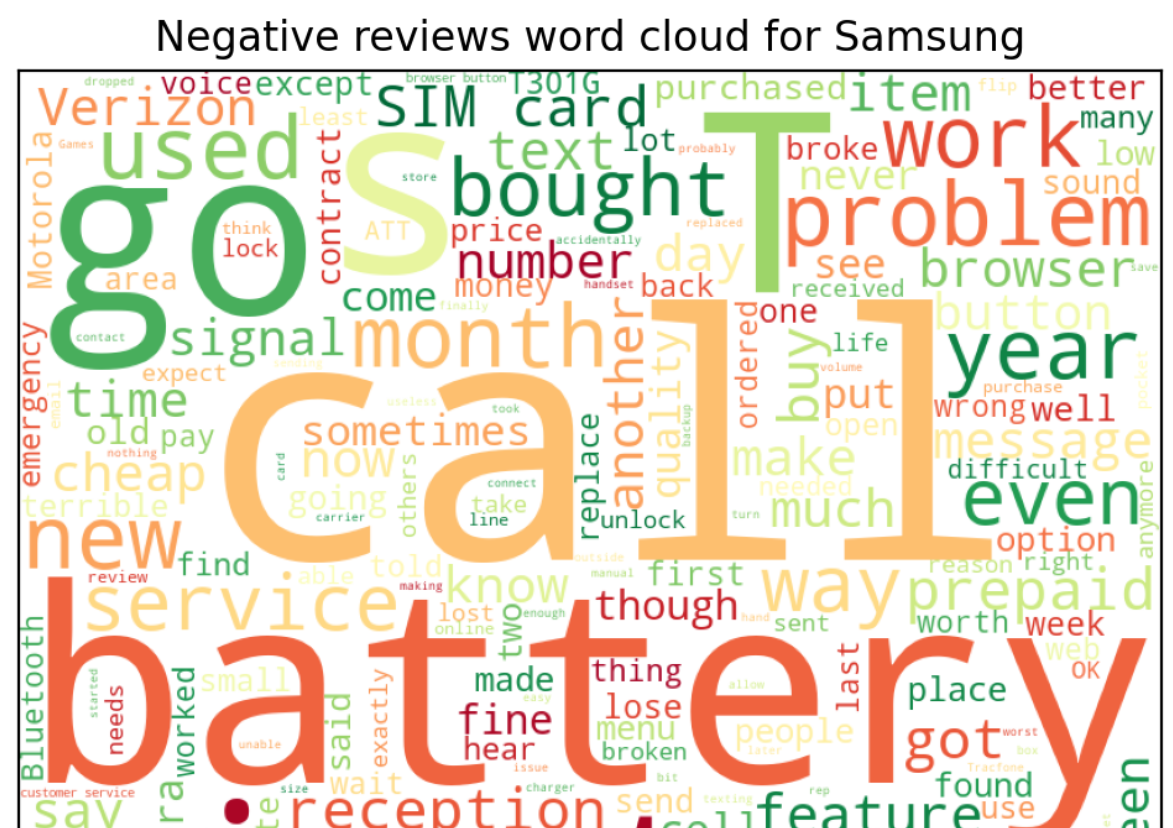
This figure describes the sentiment and trend analysis of Nokia brand. It uses Line charts and pie charts to show visualizations.



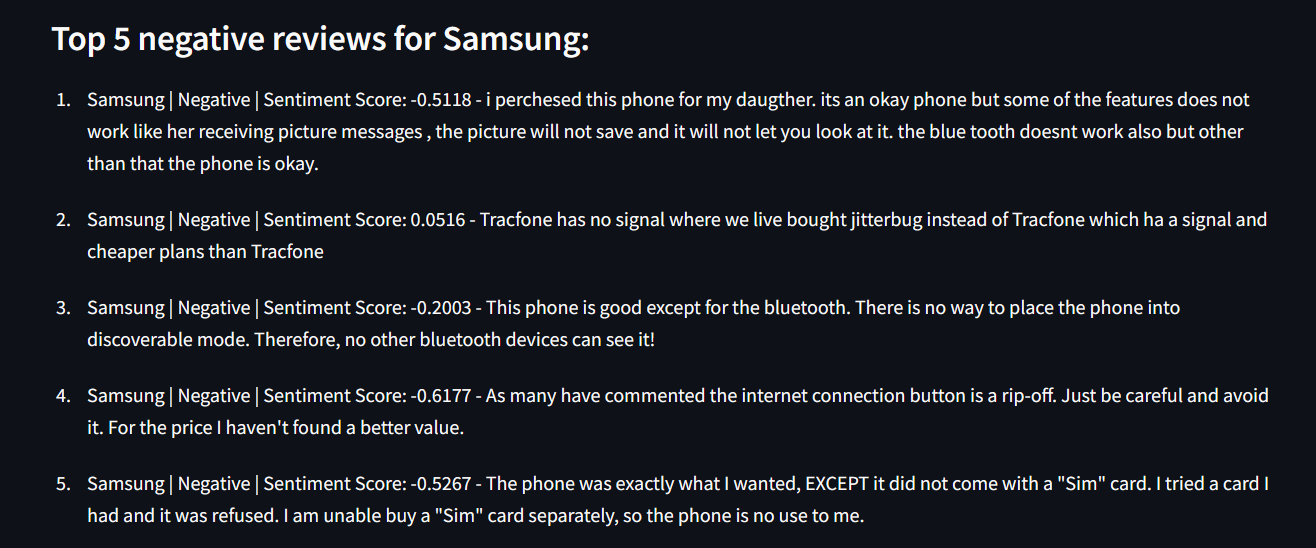
The Brand Distribution and count is visualized by using histograms.



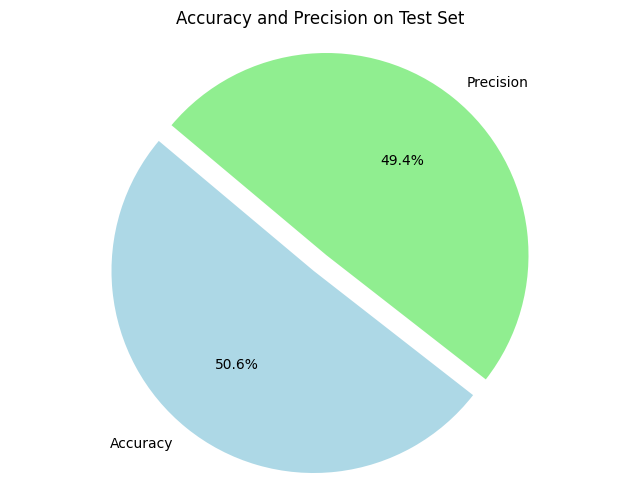
The word cloud contains all the positive words of a particular brand.



The word cloud contains all the negative words of a brand.

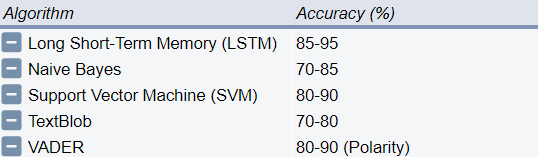


The top 5 negative reviews are given based on the frequency in order to corrent and improve the brand performance.



The VADER approach gives 50.6% accuracy and 49.4%precision where the accuracy is slightly higher than the precision so that the model performs well on the test data.

**5.1 COMPARISON WITH REGARDS TO POLARITY**



The VADER has a polarity score of 80-90 and an accuracy of 96% polarity of 80 which performs better than TextBlob [[15]](https://paperpile.com/c/ISzt1k/Ft1E) and Naive Bayes.

**References**

[1] [S. Burns, *Natural Language Processing: A Quick Introduction to NLP with Python and NLTK*. 2019.](http://paperpile.com/b/ISzt1k/UUkK)

[2] [K. Bai and K. H. Tan, “The Influence of Online Social and Physical Presence on User Consumption Decisions in TikTok Livestreaming: A Scoping Review,” *Cyberpsychol. Behav. Soc. Netw.*, May 2024, doi:](http://paperpile.com/b/ISzt1k/Ovq8) [10.1089/cyber.2023.0526.](http://dx.doi.org/10.1089/cyber.2023.0526.)

[3] [S. Raghavendra, *Beginner’s Guide to Streamlit with Python: Build Web-Based Data and Machine Learning Applications*. 2023.](http://paperpile.com/b/ISzt1k/AQyJ)

[4] [V. Vought, R. Vought, A. S. Lee, I. Zhou, M. Garneni, and S. A. Greenstein, “Application of sentiment and word frequency analysis of physician review sites to evaluate refractive surgery care,” *Adv Ophthalmol Pract Res*, vol. 4, no. 2, pp. 78–83, Mar. 2024.](http://paperpile.com/b/ISzt1k/V3Uh)

[5] [F. Millstein, *Natural Language Processing With Python: Natural Language Processing Using NLTK*. Frank Millstein, 2020.](http://paperpile.com/b/ISzt1k/jeLd)

[6] [S. Poria, A. Hussain, and E. Cambria, *Multimodal Sentiment Analysis*. 2018.](http://paperpile.com/b/ISzt1k/GBLH)

[7] [C. Chen, B. Xu, J.-H. Yang, and M. Liu, “Sentiment Analysis of Animated Film Reviews Using Intelligent Machine Learning,” *Comput. Intell. Neurosci.*, vol. 2022, p. 8517205, Jul. 2022.](http://paperpile.com/b/ISzt1k/NbPO)

[8] [S. Bhattacharyya, V. Snasel, A. E. Hassanien, S. Saha, and B. K. Tripathy, *Deep Learning: Research and Applications*. Walter de Gruyter GmbH & Co KG, 2020.](http://paperpile.com/b/ISzt1k/Y4CM)

[9] [Christodoulakis, Christina, et al. "Pytheas: pattern-based table discovery in CSV files." *Proceedings of the VLDB Endowment* 13.12 (2020): 2075-2089.](https://www.researchgate.net/publication/344969402_Pytheas_pattern-based_table_discovery_in_CSV_files)

[10] [I. C. Obagbuwa, S. Danster, and O. C. Chibaya, “Supervised machine learning models for depression sentiment analysis,” *Front Artif Intell*, vol. 6, p. 1230649, Jul. 2023.](http://paperpile.com/b/ISzt1k/n04i)

[11] [M. Falter *et al.*, “Using natural language processing for automated classification of disease and to identify misclassified ICD codes in cardiac disease,” *Eur Heart J Digit Health*, vol. 5, no. 3, pp. 229–234, May 2024.](http://paperpile.com/b/ISzt1k/ahlb)

[12] [R. Li, J. Wang, J. Li, and M. Kou, “Long, short, and medium terms wind speed prediction model based on LSTM optimized by improved moth flame optimization algorithm,” *Environ. Sci. Pollut. Res. Int.*, May 2024, doi:](http://paperpile.com/b/ISzt1k/cg0d) [10.1007/s11356-024-33580-8.](http://dx.doi.org/10.1007/s11356-024-33580-8.)

[13] [S. V. Balshetwar, A. Rs, and D. J. R, “Fake news detection in social media based on sentiment analysis using classifier techniques,” *Multimed. Tools Appl.*, pp. 1–31, Mar. 2023.](http://paperpile.com/b/ISzt1k/lJF5)

[14] [A. Schroeder, C. Mayer, and A. M. Ward, *The Book of Dash: Build Dashboards with Python and Plotly*. No Starch Press, 2022.](http://paperpile.com/b/ISzt1k/fZSr)

[15] [V. S. Dsouza *et al.*, “A sentiment and content analysis of tweets on monkeypox stigma among the LGBTQ+ community: A cue to risk communication plan,” *Dialogues Health*, vol. 2, p. 100095, Dec. 2023.](http://paperpile.com/b/ISzt1k/Ft1E)