A Comparative study on sentimental analysis of messages in a WhatsApp group using VADER , textblob and hybrid ensemble model

1.Reshma Attavara
Lecturer in Computer Science
Government Polytechnic Channapatna, Ramanagara
reshmaattavar@gmail.com

Abstract

The social landscape changes on a regular basis due to the rapid growth in the usage of social media platforms like Facebook, Instagram and X. WhatsApp has become an integral part of our daily lives. WhatsApp is a messaging application that facilitates the transmission of text messages, audio messages, video messages and other compatible files. Schools and colleges are increasingly using WhatsApp as a primary communication tool. The purpose of this study is to analyse WhatsApp group chats among students to evaluate the emotional messages sent and received by college students. Machine learning ensemble algorithms are more suitable for analysing such kind of data when compared to lexicon based VADER (Valence Aware Dictionary and sentiment Reasoner) and textblob. In the present paper, we suggest a ensemble technique that is based on Random forest with Adaboost and SVM with Adaboost using a pre-trained embedding approach. This ensemble method learns to automatically extract features for sentiment analysis and classifies WhatsApp messages or chats into two polarity of positive or negative. The model we propose to implement the result provides better performance on the benchmark dataset. Compared to the baseline machine learning methods, the performance of the ensemble method based on Adaboost has improved.

Keywords: **AdaBoost algorithm, VADER, textblob, Support Vector Machine, Random Forest ,Term frequency –Inverse Document Frequency (Tf-Idf).**

Introduction

1. witter Sentiment Analysis means, using advanced text mining techniques to investigate the sentiment of □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
2. the text (here, tweet) within the sort of positive, negative, and neutral. it's also called Opinion Mining, is □
3. primarily for analyzi

In today's digital panorama, microblogging platforms such as X, Facebook, Tumblr and WhatsApp have evolved beyond mere hubs for social interaction, transforming into treasured sources of information. Among these platforms, WhatsApp stands out as the primary arena where millions of users engage in daily exchanges through messages [1]. These group chats wield considerable influence, shaping public opinion and sentiment on a wide array of subjects, including products, topics, events, and governmental issues. These messages are monitored and analysed to provide vital insights to individuals, private entities, and public institutions [2]. Given the immense volume of data generated, sentiment analysis emerges as a vital technique for extracting actionable intelligence. By employing sentiment analysis, users can discern the prevailing sentiments expressed in tweets, thus avoiding the impracticality of manual scrutiny of millions of individual messages [2] [3] .

As a subdivision of Natural Language Processing (NLP), sentiment analysis offers a orderly approach to categorize texts into positive, negative, or neutral sentiments, facilitating efficient data analysis and decision-making processes.

Sentiment Analysis, also known as opinion mining, is a computational process [4] aimed at identifying and categorizing the various opinions expressed within a text. Its primary goal is to discern individuals' interests and attitudes toward a specific topic by analyzing the sentiment conveyed in their written expressions. Through Sentiment Analysis, algorithms analyse textual data to determine whether the expressed opinions are positive, negative, or neutral, providing valuable insights into the sentiments of individuals or groups regarding the subject matter. This process enables organizations to gauge public perception, evaluate customer feedback, and make data-driven decisions based on the sentiment trends identified within the text.

This work emphasis mainly on analysing the emotional content of user reviews and categorizing them into two main categories: positive, negative and neutral sentiments. The dataset utilized for this analysis is sourced from the college women's hostel WhatsApp group, a prominent platform for sharing the emotions and views on particular topic.

The initial phase of this research involves pre-processing the raw data to ensure its cleanliness and readiness for analysis. Following this, a variety of Natural Language Processing (NLP) techniques are employed, including tf-idf, word2vec, and countvectorizer, to transform the textual content into numerical vectors. These techniques play a crucial role in capturing the semantic meaning and context of the words within the reviews.

Once the word embedding process is complete, the transformed data is then inputted into several hybrid ensemble classification algorithms. These algorithms include AdaBoost using Support Vector Machines (SVM), Adaboost using Random Forest. The objective is to identify the most effective approach for predicting the sentiment and comparing the results with VADER and textblob.

By leveraging a combination of preprocessing, NLP techniques, and ensemble classification algorithms, this study aims to provide insights into the sentiment expressed by students in the group which helps to analyse their emotions during various phases in their college life.

Literature Survey

Sentiment analysis, often referred to as opinion mining, constitutes a branch of research dedicated to extracting and comprehending opinions from textual data. Within this field, a prevalent application involves analyzing reviews sourced from the Internet Movie Database (IMDb). Over recent years, sentiment analysis of IMDb reviews has garnered considerable interest within the research community. This attention is warranted by the potential insights that analyzing IMDb reviews can offer into the sentiments and attitudes of viewers towards movies [5].

By scrutinizing these reviews, researchers can discern trends, preferences, and criticisms voiced by audiences, thereby enriching our understanding of how films are perceived and received by the public. Such studies can inform filmmakers, critics, and industry professionals

in tailoring their productions, marketing strategies, and creative endeavours to align with audience expectations and preferences.

Classifying the sentiment expressed in a given text as positive, negative, or neutral poses one of the primary challenges in sentiment analysis. To tackle this challenge, researchers have proposed and explored various approaches in the literature. One common strategy involves leveraging machine learning techniques, encompassing both supervised and unsupervised methods [6].

In supervised learning, models are trained on labelled datasets, where each text sample is associated with its corresponding sentiment label (e.g., positive, negative, or neutral). Popular supervised learning algorithms used for sentiment classification include Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and Decision Trees and ensemble techniques. These algorithms learn to classify text based on features extracted from the text, such as word frequencies, n-grams, or word embeddings.

Other approaches in sentiment analysis include lexicon-based methods, where sentiment scores are assigned to words or phrases based on pre-defined sentiment lexicons. These scores are then aggregated to determine the overall sentiment of a text. Lexicon-based methods are particularly useful for languages or domains where labeled data is scarce.

The study [7] makes an effort to employ TextBlob and the VADER sentiment analyser in tandem with historical tweets to analyze emotional responses to the coronavirus pandemic (COVID-19). With the global impact of COVID-19, understanding public sentiment towards the pandemic is crucial for various stakeholders, including policymakers, health organizations, and the general public.

The study [8] evaluates the reliability of several hybrid approaches across a diverse range of datasets. By comparing hybrid models to individual models, the author assesses their performance across different domains and datasets. The analysis encompasses text data from tweets and reviews, and employ deep learning systems for sentiment analysis.

It's common knowledge that conventional messages [9] , like reviews of movies or products, still dominate tweet sentiment analysis, despite notable advancements as deep learning becomes more widely used and large data sets—rather than just emoticons and hashtags—become available for training. However, previous studies on opinion analysis have mostly focused on tweets, meaning that only two types of global polarities—positive and negative—occur with their work/validation/test data sets.

Proposed Methodology

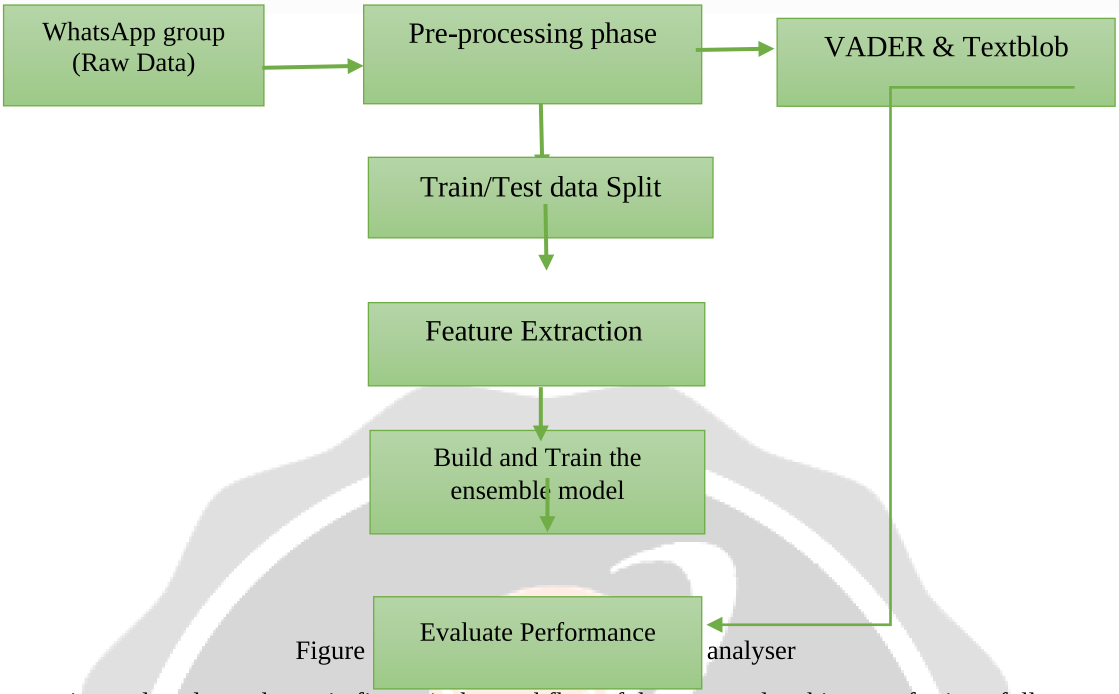


Figure analyser

For the sentimental analyser shown in figure 1, the workflow of the proposed architecture for is as follows

1. Input Dataset: The process begins with the input dataset containing raw data to be analyzed for sentiment. This dataset likely includes textual data that is exported from students whatsapp chatgroup which consists of reviews, comments, emotions in social media posts.

2. Pre-processing Stage: The raw data undergoes pre-processing to clean and transform it into a format suitable for analysis. Pre-processing techniques may include text normalization, removing stopwords, handling punctuation, and stemming or lemmatization.

3. VADER and TextBlob Analysis:Processed data is passed through VADER and TextBlob for sentiment analysis. Both VADER and TextBlob generate results for positive, negative, and neutral sentiments.

4. Machine Learning Classifiers: The labeled dataset is split into training and testing datasets.Feature extraction techniques are applied to the data to prepare it for input into different hybrid ensemble techniques.

5. Feature Extraction: Feature extraction techniques are applied to convert the textual data into numerical representations suitable for machine learning algorithms. Common feature extraction methods include TF-IDF, word embeddings (e.g., Word2Vec), and Bag-of-Words representations.

6.Hybrid Ensemble Techniques:

The pre-processed and feature-extracted dataset is fed into different hybrid ensemble techniques. Specifically mentioned are Adaboost algorithm with SVM and Adaboost algorithm with Random Forest.

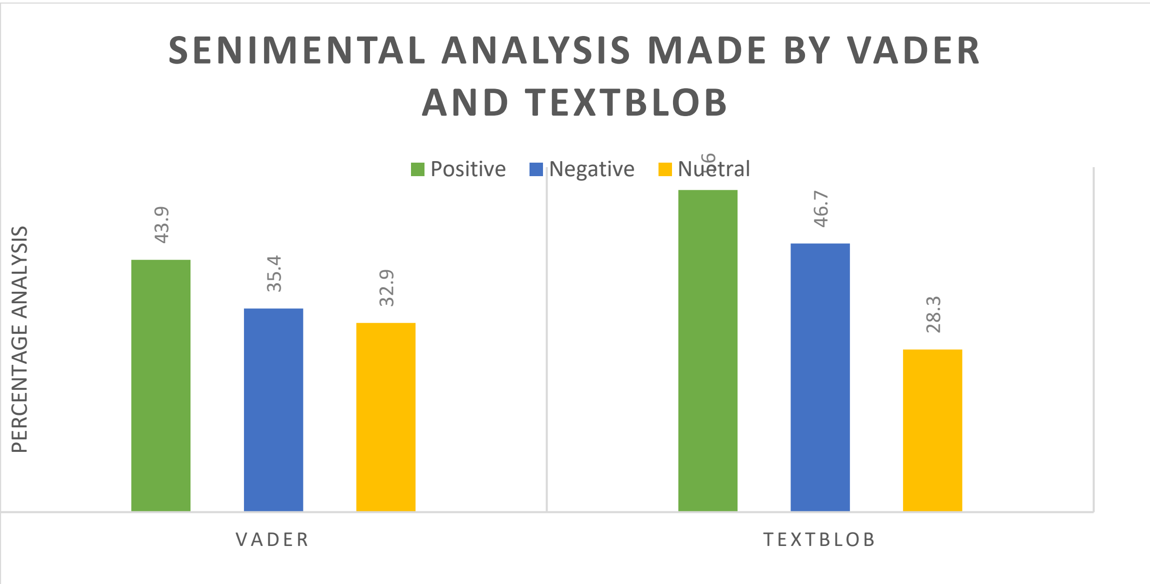
7. Model Evaluation:

The performance of each model is evaluated using a confusion matrix. Evaluation metrics such as accuracy, precision, recall, and F1-score are likely calculated to assess the effectiveness of the models.

This workflow shown in figure 1 outlines a structured approach to sentiment analysis, incorporating both rule-based (VADER and TextBlob) and ensemble machine learning-based techniques. It allows for the comparison of different methods and evaluates their performance in classifying sentiment in textual data.

Results and Discussion

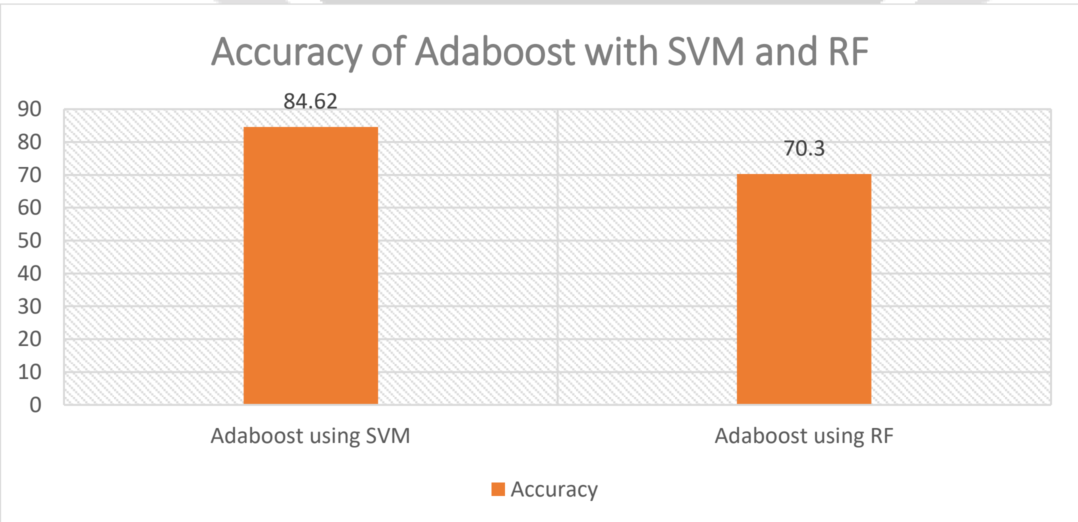
Based on the experimental results:



- VADER sentiment analysis returned 43.9% positive, 35.4% neutral, and 32.9% negative sentiment.

- TextBlob sentiment analysis returned 56.0% neutral, 46.7% positive, and 28.3% negative sentiment.

Furthermore, the performance of different hybrid ensemble techniques using AdaBoost algorithm with SVM and AdaBoost algorithm with Random Forest (RF) for sentiment analysis is evaluated:



AdaBoost algorithm with SVM achieved the highest accuracy of 84.62%. AdaBoost algorithm with Random Forest achieved an accuracy of 70.3%.The least performance is observed with AdaBoost algorithm with Random Forest.

These results indicate that AdaBoost algorithm with SVM outperforms AdaBoost algorithm with Random Forest in terms of accuracy for sentiment analysis. It's important to note that the choice of algorithm and model configuration can significantly impact the performance of sentiment analysis models. Further analysis and experimentation may be necessary to optimize the sentiment analysis pipeline and improve accuracy.

Conclusion

It can be observed that VADER and TextBlob sentiment analysis provided insights into the sentiment distribution within the dataset, with some variations in the results. VADER identified more positive sentiment compared to TextBlob, while TextBlob assigned a higher percentage of neutral sentiment. AdaBoost algorithm with SVM achieved the highest accuracy of 84.62% for sentiment analysis. AdaBoost algorithm with Random Forest exhibited lower performance with an accuracy of 70.3%.This work can be improved by using word embeddings or deep learning architectures for feature representation. Investigate other ensemble learning approaches beyond AdaBoost, such as Gradient Boosting Machines (GBM) or Stacking, to potentially improve accuracy and robustness. Utilize domain-specific sentiment lexicons or dictionaries tailored to the context of the dataset to enhance the accuracy of sentiment analysis. Explore the use of deep learning models, such as Recurrent Neural Networks (RNNs) or Transformer-based architectures (e.g., BERT), for sentiment analysis, which have shown promising results in capturing complex patterns in textual data.

By implementing these future enhancements, it is possible to further improve the accuracy, efficiency, and applicability of sentiment analysis techniques for various domains and applications. Additionally, conducting thorough evaluations and validations will ensure the reliability and effectiveness of the sentiment analysis models in real-world scenarios.

References

[1]	Cruz, E. G., & Harindranath, R. (2020). WhatsApp as ’technology of life’: Reframing research agendas. First Monday, 25(12). https://doi.org/10.5210/fm.v25i12.10405
[2]	Setting the future of digital and social media marketing research: Perspectives and research propositions, International Journal of Information Management,Volume 59,2021,102168, ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2020.102168
[3]	International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2022, Vol. 18, Issue 3, pp. 147-158 Students’ Usage of WhatsApp Instant Messenger as a Supporting Tool for Learning in Kaduna State, Nigeria
[4]	Zhang, L., Liu, B. (2016). Sentiment Analysis and Opinion Mining. In: Sammut, C., Webb, G. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7502-7_907-1
[5]	Predicting Users’ Movie Preference and Rating Behaviour from Personality and Values, Euna Mehnaz Khan
[6]	Zhang, L., Liu, B. (2016). Sentiment Analysis and Opinion Mining. In: Sammut, C., Webb, G. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7502-7_907-1
[7]	Abiola, O., Abayomi-Alli, A., Tale, O.A. et al. Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. Journal of Electrical Systems and Inf Technol 10, 5 (2023). https://doi.org/10.1186/s43067-023-00070-9
[8]	Mohammed Hasan Ali Al-Abyadh, Mohamed A. M. Iesa, Hani Abdel Hafeez Abdel Azeem, Devesh Pratap Singh, Pardeep Kumar, Mohamed Abdulamir, Asadullah Jalali, "Deep Sentiment Analysis of Twitter Data Using a Hybrid Ghost Convolution Neural Network Model", Computational Intelligence and Neuroscience, vol. 2022, Article ID 6595799, 8 pages, 2022. https://doi.org/10.1155/2022/6595799
[9]	Neha et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012114