Analysis of Emotions made in Instagram posts using XGBoost and AdaBoost algorithm

1. **Reshma Attavara ,** *Senior Grade Lecturer in Computer Science, Government Polytechnic Channapatna, Ramanagara,* reshmaattavar@gmail.com
2. **Archana B S,** *Senior Grade Lecturer in Computer Science, Government Polytechnic Channapatna, Ramanagara,*
3. **Rashmi T N,** Lecturer *in Computer Science, Government Polytechnic Channapatna, Ramanagara,*

Abstract

Finding the emotional tone or ideas expressed in text is the goal of emotional analysis, also known as opinion mining. When we use in social media sites such as Facebook, Instagram, X and WhatsApp, it helps in examining user-generated content to determine the general audience attitude, which may be anything from positive or negative emotions. This procedure aids in illuminating the sentiment behind every post or remark. In this paper we use hybrid machine learning technique to determine the sentiment behind every comments made by the user in the Instagram site.

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

Introduction

In this current digital environment, social media sites such as X, Instagram, Facebook and WhatsApp are used by millions of people for the purpose of expressing their feelings or opinion about society, and various other topics that occur around them. These social media participants may use critical remarks as a way to release their feelings that have been accumulated within them using positive or negative comments.
One particular focus of this research involves sentiment analysis, which seeks to analyse and comprehend the sentiment or opinion conveyed in a given piece of text [1].
The [2] paper conducts a comprehensive literature review of Instagram, with a particular focus on the research methodologies employed by researchers to gather and analyse data.
The chats exert significant influence, moulding public opinion and sentiment on a variety of subjects such as products, topics, events, and governmental issues. These messages are monitored and analysed to offer crucial insights of individuals about private entities, government, political disputes and public institutions [3].

This study primarily focuses on analysing the emotions expressed by users and categorizing them into two main categories: positive and negative. The dataset utilized is from the kaggle website, a popular platform for collecting the dataset. Initially, the raw data undergoes pre-processing to ensure it is clean and ready for analysis. Subsequently, various Natural Language Processing (NLP) techniques such as tf-idf, word2vec, and countvectorizer are employed to convert the text into word embeddings. These embeddings are then used as input into several hybrid ensemble classification algorithms including AdaBoost with SVM and XG Boost. The goal is to determine the most effective approach for accurately predicting emotions of Social media participants.

Literature Survey

The research [4] aimed to assess customer satisfaction with digital payment services (Go-Pay, Ovo, and LinkAja) in Indonesia, using Instagram sentiment analysis based on textual data extracted from Instagram comments. The sentiment analysis classification algorithms Naïve Bayes and K-Nearest Neighbors (KNN), along with customer satisfaction theory, were used to build the research model.

One of the main problems with sentiment analysis is determining if a text's sentiment is good, negative, or neutral. In the literature, scholars have put up and examined a number of strategies to address this problem. Using machine learning techniques—which include both supervised and unsupervised methods—is one popular tactic [5]. Models are trained using labeled datasets in supervised learning, where each text sample has a matching sentiment label (e.g., positive, negative, or neutral). The supervised learning methods Naive Bayes, Decision Trees, Logistic Regression, Support Vector Machines (SVM), and ensemble approaches are commonly employed for sentiment classification. Based on characteristics that are taken from the text, such as word frequencies, n-grams, or word embeddings, these algorithms learn to categorize text.

In the work [6], sentimental categories are offered as a way to categorize the feelings. Sentimental keywords for important sentiments are extracted using hashtags, which are integral parts of Instagram. Sentiment categories may be applied to user postings, and sentiments can be discovered by measuring the similarity between the sentiment keywords and the sentiment adjective candidates.

The project [7] aims to investigate emotional reactions to the coronavirus pandemic (COVID-19) by utilizing TextBlob and the VADER sentiment analyzer in conjunction with historical tweets. Given the worldwide reach of COVID-19, it is imperative for politicians, health organizations, and the general public to comprehend public opinion around the pandemic.

The article describes the usage of a unique cluster-based classification model for online product reviews by P. Vijayaragavan et al. [8]. The model being displayed is made up of several processes. Initially, an SVM-based classification model is used to categorize the product reviews. After that, a confusion matrix is constructed to take into consideration the probability that every client would purchase the product. Next, the K-means clustering technique divides the given data into two groups.

The study [9] looked at the best ways to use hashtags (#) in instagrammable posts to communicate ideas. More precisely, we looked at the contents and usage of 'Card News,' which is a series of images that tell a short tale. Because they often convey the post's intended audience for sharing, hashtags are essential. According to our research, one can learn a lot about individuals' Instagram activity by looking at how hashtags are used in Card News postings.
.

Implementation:

• **Data Collection**: Gathering Instagram posts and their associated comments or captions from kaggle website. This data is used as the basis for emotional analysis.

• **Data Preprocessing**: Clean and pre-process the textual data. This involves:

- Removing special characters, emojis, and non-alphanumeric characters.
- Tokenizing the text into words or phrases.
- Removing stop words (commonly used words that do not contribute to the meaning).
- Stemming or lemmatizing words to reduce them to their root form.

• **Feature Extraction**: Convert the pre-processed text into numerical features that can be used by machine learning algorithms. Common techniques include:

- TF-IDF (Term Frequency-Inverse Document Frequency): Measures how important a word is to a document within a collection or corpus.
- Word embeddings (such as word2vec): Represents words as dense vectors of real numbers.

• **Training and Testing Data**: Split the dataset into training and testing sets. Typically, the training set (about 70-80% of the data) is used to train the model, while the testing set (about 20-30% of the data) is used to evaluate its performance.

• **Model Selection**: Choose AdaBoost and XGBoost as the classification algorithms for emotional analysis. AdaBoost combines multiple weak classifiers to create a strong classifier, while XGBoost is an optimized gradient boosting algorithm known for its speed and performance. In this work we used SVM as a weak classifier to create an Adaboost model.

• **Model Training**: Train the AdaBoost and XGBoost models are used on the training dataset. During training, the models will learn to classify the text based on the emotional sentiment expressed in the Instagram posts.

• **Model Evaluation**: Evaluate the performance of the trained models using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. This step helps to assess how well the models can predict the emotional sentiment of Instagram posts.

Results and Discussion

The model evaluation and the results of 2 hybrid machine learning models is compared using the graph shown in **figure 1**. Additionally, the effectiveness of 2 hybrid ensemble methods for emotional analysis is calculated. Based on its classification report comparison is made.

Table 1: Comparison table of performance measure given by different classification model

Models used	Accuracy	F1-Score	Precision	Recall
AdaBoost using SVM	87	90.1	90	90
XGBoost Algorithm	90.3	94.7	94	93

When we observe Table 1, it can be clearly noted that the performance of XGboost Algorithm is better than AdaBoost model. XGBoost has scored 90.3% accuracy whereas AdaBoost has scored 87% accuracy.

The accuracy, precision, recall and F1-scores for the applied classifiers have been summarised. Meanwhile, the F1score of both the models has very little difference when comparison is made.

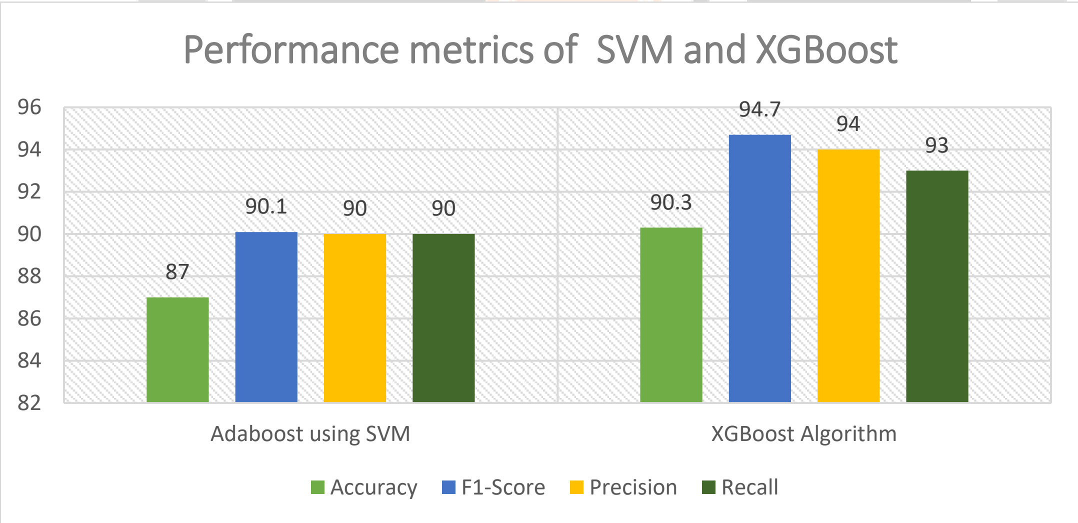

Figure 1: Comparison of AdaBoost and XGBoost Algorithms .

Conclusion

When it comes to analysing emotions, both Adaboost and XGBoost algorithms prove to be effective tools, each with distinct strengths. XGBoost achieves a high accuracy of 90%, showcasing its robustness in capturing intricate emotional patterns and making it a reliable choice for applications requiring precise emotion detection. Adaboost, with its 87% accuracy, also demonstrates commendable performance, particularly in scenarios where sequential adjustment of weights can help prioritize complex emotional nuances. Ultimately, the choice between these algorithms depends on specific needs such as computational efficiency, interpretability, or the emphasis on handling difficult cases in emotional analysis. Both Adaboost and XGBoost offer valuable contributions for advancing emotion analytics, each contributing uniquely to the field's development.

References

[1]	Su Y, Yuan D, Chen DG, Ng RH, Wang K, Choi J, Li S, Hong S, Zhang R, Xie J, Kornilov SA, Scherler K, Pavlovitch-Bedzyk AJ, Dong S, Lausted C, Lee I, Fallen S, Dai CL, Baloni P, Smith B, Duvvuri VR, Anderson KG, Li J, Yang F, Duncombe CJ, McCulloch DJ, Rostomily C, Troisch P, Zhou J, Mackay S, DeGottardi Q, May DH, Taniguchi R, Gittelman RM, Klinger M, Snyder TM,. Multiple early factors anticipate post-acute COVID-19 sequelae. Cell. 2022 Mar 3;185(5):881-895.e20. doi: 10.1016/j.cell.2022.01.014. Epub 2022 Jan 25. PMID: 35216672; PMCID: PMC8786632.
[2]	Yang, Chen. (2021). Research in the Instagram Context: Approaches and Methods. The Journal of Social Sciences Research. 7. 15-21. 10.32861/jssr.71.15.21.
[3]	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
[4]	H. Sudira, A. L. Diar and Y. Ruldeviyani, "Instagram Sentiment Analysis with Naive Bayes and KNN: Exploring Customer Satisfaction of Digital Payment Services in Indonesia," 2019 International Workshop on Big Data and Information Security (IWBIS), Bali, Indonesia, 2019, pp. 21-26, doi: 10.1109/IWBIS.2019.8935700. keywords: {Sentiment analysis;Customer satisfaction;Classification algorithms;Text mining;Customer Satisfaction;Digital Payment;Instagram;K-Nearest Neighbors;Naïve Bayes;Sentiment Analysis},
[5]	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
[6]	Gangrade, Shweta and Shrivastava, Nirvishesh and Gangrade, Jayesh, Instagram Sentiment Analysis: Opinion Mining (April 16, 2019). Proceedings of Recent Advances in Interdisciplinary Trends in Engineering & Applications (RAITEA) 2019,Available at SSRN: https://ssrn.com/abstract=3372757 or http://dx.doi.org/10.2139/ssrn.3372757
[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]	P. Vijayaragavan, R. Ponnusamy, and M. Aramudhan, “An optimal support vector machine based classification model for sentimental analysis of online product reviews”, Future Generation Computer Systems, vol. 111, 2020, 234–240.
[9]	M J Nam , J I Kim , J H Shin .A User Emotion Information Measurement using Image and Text on Instagram Based Journal of Korea Multimedia Society , volume 17 , issue 9 , p. 1125 - 1133 Posted: 2014