A Factual Sentiment Analysis on Instagram Data – A Comparative Study Using Machine Learning Algorithms.

Amrutha Ramachandran
Department of Computer Science & IT
Amrita School of Arts and Sciences, Kochi, India
amrur782@gmail.com

Swetha Ashok
Department of Computer Science & IT, Amrita School of Arts and Sciences, Kochi, India
swethaashokmini1999@gmail.com

Remya Nair T
Department of Computer Science & IT
Amrita School of Arts and Sciences, Kochi, India
remybhi@gmail.com

2022 Algorithms, Computing and Mathematics Conference (ACM) | 978-1-6654-9252-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACM57404.2022.00009

***Abstract*— Social media is one of the most significant parts of our daily life. Our social media profiles are a reflection of our emotions. Instagram is the world's most popular photo-based social networking platform, with a reasonably high number of users ranging from regular people to artists, public figures, and top authorities. Users on Instagram may add captions to their images to make them more interesting. In this study, we are focusing on conducting sentiment analysis on Instagram captions by applying three different algorithms. We are concluding that the Logistic Regression algorithm is outperforming along with SMOTE and VADER compared to XG Boost and Random Forest algorithms. We started by acquiring data and dividing it down into little tokens, then we remove connection words and give clean data via the stop word removal mechanism. The cleaned data is then passed via the NLTK (Natural Language Toolkit) passer, which uses the VADER sentiment unit to produce sentiment based on the data. Then applying different algorithms XGBoost, Logistic Regression, and Random Forest on the produced sentiment. The accuracy of algorithms such as XGBoost, Logistic Regression, and Random Forest on sentiment data was also analyzed and tested and can be concluded that Logistic Regression performed well on these kinds of data with more accuracy. Through this work, the accuracy is lifted to a better level and thereby getting a truthful idea of the Instagram captions.**

Keywords*— *Sentiment Analysis, XGBoost, Logistic Regression, Accuracy

I. INTRODUCTION

Sentiment analyzation is the process of recognizing and segmenting the views and attitudes represented in a source text. Instagram is a social media platform for sharing photos, videos, and images. Instagram has a relatively large number of users from all fields of life, ranging from ordinary people to artists, public personalities, and senior officials, making it the world's most popular photo-based social networking platform. Instagram users may add captions to their photos, giving them more life.

In this paper we would like to showthe sentiment analysis in Instagram caption by using three different algorithms. Firstly, we examine the caption and divide it into two polarities (Positive, Negative). Positive emotion is indicated by a polarity value of 0; negative sentiment is indicated by a polarity score of 1.Then, using algorithms, we train the data and predict sentiment. The algorithms used in this work are XGBoost, Logistic Regression, and Random Forest.

II. LITERATURE REVIEW

The authors used Twitter to track and predict depression in people. To characterize depressive behavior, they used crowd sourcing and presented a number of social media indicators such as emotions, language, and user interaction. They also discovered that people with depression had less social activity, more negative feelings, higher self-attention concentration, and more relational and medical concerns. They created an SVM classifier with different properties which can identify an individual's risk of developing depression before the onset of depression is observed [1]. Sentimental Analysis of Facebook

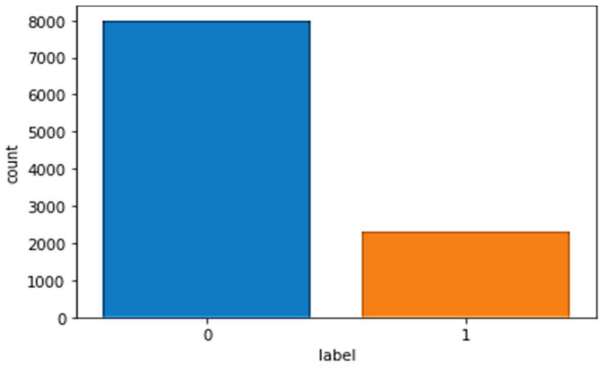
978-1-6654-9252-2/22/$31.00 ©2022 IEEE
DOI 10.1109/ACM57404.2022.00009

was used for this project. To extract data regarding a user's sentiment which could be positive, neutral, or negative as expressed in the message. The text messages are analyzed, and just the messages that are required are filtered and classified into emotions. Every week, the emotional shifts will be noted. The methods used for classification are lexical-based and machine-learning techniques [2]. To classify an individual's emotional transformation, the study used polarities (positive, neutral, and negative). They developed an app that extracts comments, messages, and likes from profiles, categorizes communications by polarity, and develops and maintains the user sentiment profile. On a weekly basis, the emails will be evaluated, with the essential messages filtered or classified. Lexical analysis and machine learning methodologies were applied in the application's classification process [3]. In this study uses ad hoc analysis to analyze sentiment on captions on Instagram using hash tags (#read, #reading) from modern readers. They can gain a better understanding of existing Instagram readers' interests through this analysis. Public libraries could use the polarities and feelings stated about the topic to guide their future strategies. Positive (Love, Joy) and Negative (Sadness, Fear, Anger) emotions are the most common classifications (Surprise). They show how user-generated content on Instagram can provide librarians with useful information. Analyses can help improve library services by identifying and interpreting patterns [4]. The text's nature has been characterized as positive, negative, or neutral by Lattha A and Hemanh Kumar. For identifying the text data gathered from twitter feeds, they employed Naive Bayes and a support vector machine technique. For doing emotional analysis on depression, I used natural language processing on Twitter feeds. Removal of emoji, hyperlinks, slang substitution, timestamps, digits, spelling correction, proper nouns, and tokenizing are some of the methods used in this research [5]. To assess emotional transformation, two sorts of polarities (positive and negative) are used in this study. Sentiment analysis, subjectivity analysis, or polarity calculations are the methods employed. They analyzed sentiment in text communications using supervised machine learning algorithms, such as online product evaluations, public tweets on Twitter, and film reviews. Messages are pre-processed until being examined using three distinct machine learning approaches for sentiment analysis: Naive Bayes, Decision Trees, and SVM[6].SimranPatil, Arif Shaikh, and Sakshi Singh developed a sentiment analysis model that allows users to categorize words based on their sentiments, such as whether they are pleasant or negative, as well as the amplitude of their feelings. Sentiment analysis models concentrate on polarity, sentiments and emotions, urgency, and even intentionality [7]. In this work, they used sentiment analysis to determine the user's positive emotions (such as happiness, pleasure, surprise, and excitement) as well as negative emotions (such anxiety, stress, sadness, anger). Negative feelings will then be classified, and the severity of depression will be displayed. The purpose of the study was to keep track of the users' feelings [8]. In this work, they have assessed people's opinions and beliefs on many topics .NLP and machine learning are used in this. The purpose of this study is to determine review orientation automatically. To address this issue, a hybrid solution is offered [9]. This research project is focused on the sentiment analysis of Amazon data from product reviews utilizing the Twitter API. SVM and KNN comparison study was done by Kajal and Prince Verma. Due to KNN's usage of several hyper planes for data classification, it outperforms SVM classification in terms of performance. Both strategies are implemented in Python, and analysis of the experiment's findings reveals that the KNN strategy outperformed the SVM approach in terms of accuracy. In comparison to SVM, the KNN technique has a shorter execution time [10]. Dr. R. Ramachandiran , Arunnkumar and Balachande uses a number of strategies to focus on Twitter sentiment analysis. The tweets will first be converted into structured format, and then the tweets are resolved utilizing libraries that use the Twitter API. Algorithms must be used to train the database so that it can verify tweets and extract the necessary sentiments from the feed. The precise user review on a particular product is predicted and identified by machine learning algorithms [11]. In order to forecast suicidal behaviors supported by the severity of depression, Prof. S.J. Pachouly, GargeeRaut, KshamaBute, RushikeshTambe, ShrutiBhavsar, has presented a system for analyzing depression and detecting suicidal ideation. They trained and evaluated classifiers to find tweets from potentially sad twitter users. This classifier uses information extrapolated from user behavior in tweets to determine if a user is depressed or not. On a scale of 0 to 100%, classification machine techniques are used to train and categorize the machine in various phases of depression. Using machine learning classification algorithms, the data gathered from the tweets was divided into depressed and non-depressed tweets as part of a predictive method for the early detection of depression or mental illness [12]. With the help of various tools and frameworks, the Arpit Upadhyay, Nishi Sharma, Aryan Chaudhary and Divya Jain created an application while trying to understand sentiment analysis and opinion mining on a deeper level through their research. For file processing and sentence categorization, they used a Naive Bayes analyzer and a Pattern-Based analyzer, respectively. The polarity and subjectivity attributes that the TextBlob returns are utilized to determine the sentiment [13]. This paper provides a summary of the various levels of sentiment analysis, uses the Twitter dataset, and conducts experiments. To categorize the tweets from Twitter, Princy Sharmal and Vibhakar Mansotra used Naive Bayes, Support Vector Machine, and Random Forest. The effectiveness of these algorithms was then compared [14]. The authors of this study, Pratiksha Karpe, Rakhi Marathe, Prachi Kolekar, and Prof. Namrata Wasatkar, are attempting to determine the author of a book or novel by examining the text provided by the user and examining its language, vocabulary, and use of various words according to its tone. They have used Natural Language Processor for data pre-processing. Sentiment Analysis is advocated using Naive Bayes [15].

III. METHODOLOGY

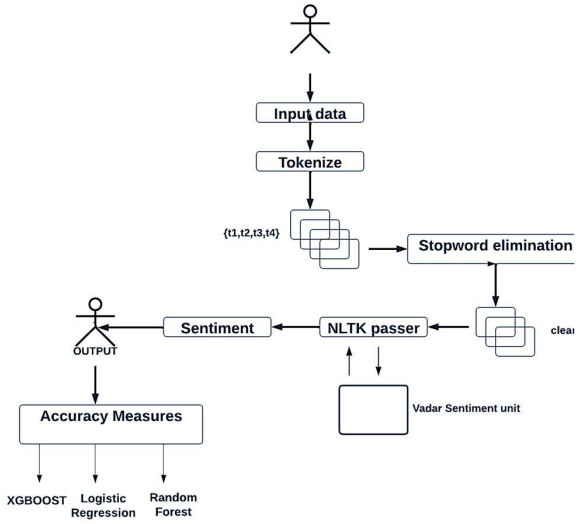
Dataset

The datasets utilized were obtained from Kaggle. It has 10314 records in three columns: Unnamed, Message, and Label. Message and label columns are the most commonly used columns**.** It has 0 as positive caption and 1 as negative polarity. The result obtained after doing data analysis in CoLab is shown below. The graph showing the count of 0 polarity score and 1 polarity score in the dataset:



Methods used

We practiced the machine learning algorithms, XGBoost, Logistic Regression, and Random Forest on the concerned data. The entire workflow is shown in the diagram given below.



We begin by gathering data and breaking it down into small tokens such as t1, t2, t3, and t4, after which we do stop word removal to remove connector words and provide clean data. The cleaned data is then sent through the NLTK (Natural Language Toolkit) passer, which generates sentiment utilizing the data using the VADER Sentiment unit. The result is then produced as an output.

The generated sentiment (output of VADER sentiment unit) calls for accuracy measures using the algorithms XGBoost, Logistic Regression, and Random Forest. The accuracy of these algorithms is then compared. The higher accuracy rate shows the best algorithm among the three algorithms.

NLTK

The Natural Language Toolkit, or NLTK for short, is used in natural language processing to work with human language data (NLP).

NLP

The systematic processing of human language by software, such as speech and text, is known as Natural Language Processing or NLP for short. NLP draws on a range of sciences, primarily computer science, to fill the gap among human interaction and machine comprehension.

Logistic Regression

A set of variables is predicted using Logistic Regression. Since only numerical data will be accepted in LR, it should be balanced. For that we applied smote analysis on the data. As a result the algorithm brought an accuracy of 0.98 for our work.

XGBoost

XGBoost is a framework that may be used in a variety of languages. It is platform independent. In the study XGBoost showed 0.95 accuracy score.

Random Forest

Random forest is a supervised learning method. It can be applied to both classification and regression. It's also the algorithm that's the most adjustable and user- friendly. In our paper, the accuracy score is 0.95.

IV. CRITICAL ANALYSIS

Imbalanced data

Data that is not uniform is considered Imbalanced data. Imbalanced data occurs when one group of classes exceeds over another. As a result, the machine learning model becomes

more biased towards the majority class. It leads to incorrect categorization of minority groups. The data's accuracy would be low if we use imbalanced data. The imbalanced data will be effectively handled by the Logistic Regression algorithm with the help of SMOTE technique. So we are able to make use of the better percentage of available data. Thereby reducing the loss of data during the data preprocessing phase. Because of high end preprocessing using SMOTE the Logistic Regression algorithm out performs compared to other algorithms XGBoost and Random Forest. We also used Imblearn library for effectively importing data.

Smote

SMOTE (synthetic minority oversampling technique) is one of the most extensively utilized oversampling techniques for resolving the imbalanced problem. Its purpose is to produce a more evenly distributed distribution of classes by randomly reproducing minority class cases. It is used to construct synthetic class samples of the minority class in order to balance the distribution, and then it is used to clean irrelevant points in the boundary between the two classes in order to improve the separation between two classes using the sampling technique.

Imblearn

Assists us in creating a data collection with an equal number of classes. This sort of data collection would allow the prediction model to generalize effectively. It aids in resampling classes that are oversampled or undersampled in the first place. When the imbalance ratio is significant, the output is skewed toward the class with the most examples.

VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is particularly suited to sentiments shared on social media. It reveals the Positivity and Negativity score as well as positive or negative sentiment. The sentimental analysis of VADER is based on a lexicon that maps lexical information to emotion intensities known as sentiment scores. A text's sentiment score may be determined by adding the intensity of each word in the text.

Experiment and Result Analysis

A. Now measuring the accuracy of various algorithms using the obtained sentiment. The processing of sentiment with different algorithms is carried out in COLAB and obtained results are shown below. The result is mainly focused on four parameters – precision, recall, f1-score and support.

XGBoost

	precision	recall	f1-score	support
0	0.95	0.95	0.95	2595
1	0.95	0.95	0.95	2595
accuracy			0.95	5190
macro avg	0.95	0.95	0.95	5190
weighted avg	0.95	0.95	0.95	5190

Logistic Regression

	precision	recall	f1-score	support
0	0.98	0.99	0.98	2595
1	0.99	0.97	0.98	2595
accuracy			0.98	5190
macro avg	0.98	0.98	0.98	5190
weighted avg	0.98	0.98	0.98	5190

Random Forest

	precision	recall	f1-score	support
0	0.95	0.95	0.95	2595
1	0.95	0.95	0.95	2595
accuracy			0.95	5190
macro avg	0.95	0.95	0.95	5190
weighted avg	0.95	0.95	0.95	5190

The above shown are the accuracy scores of three algorithms used in this paper.

TABLE 1. Represents the accuracy

Algorithm	Accuracy
XGBoost	0.95
Random Forest	0.95
Logistic Regression	0.98

V. CONCLUSION

The NLTK and the VADER analyzer were used to do sentiment analysis on INSTAGRAM captions in this study. VADER was able to classify large amounts of data rapidly and easily. The VADER sentiment output is then processed for accuracy measures using various ML classification algorithms like XGBoost, Logistic Regression, and Random Forest. The support of VADER and SMOTE push the performance of the Logistic Regression algorithm to the highest. The result indicates that Logistic Regression is the best of the three and Logical regression appears to have a greater accuracy rate than Random Forest and XGBoost. By applying SVM and lexicon tagging, it was possible to get an accuracy score of 0.83 on the base paper. The result of this study shows that by applying Logistic Regression along with the appropriate use of SMOTE and VADER, the accuracy rate is improved to 0.98. So suggesting the Logistic Regression algorithm as a better choice for analyzing the response of social media to get a truthful vision.

VI. REFERENCES

[1] Munmun De Choudhury Michael Gamon Scott Counts Eric Horvitz, “Predicting Depression via Social Media”, 2013 International AAAI Conference on Weblogs and Social Media.

[2] Alvaro Ortigosa, Jose M. Martín, Rosa M. Carro,” Sentiment analysis in Facebook and its application to e-learning”, 2013.

[3] Ming Zhan , RuiboTu ,Qin YU, ” Understanding Readers:Conducting Sentiment Analysis of Instagram Captions”, 2018International Conference on Computer Science and Artificial Intelligence.

[4] HemanhKumar ,Lattha A,” Depression detection with sentiment analysis of tweets”, 2019 (IRJET).

[5] Abhishek Bhagat , Akash Sharma , Sarat Kr. Chettri, ”Machine Learning- Based Sentiment Analysis for Text Messages”, 2020.

[6] SimranPatil, Arif Shaikh, Sakshi Singh,”Sentiment Analysis using machine Learning”, 2021 (IRJET).

[7] Princy Sharma, Prof. VibhakarMansotra,” A study of social sentiment analysis in the times of covid -19 using twitter”, (IRJET).

[8] Abhishek Chaube,Vaidehi Dani, Trupti Dhapola, Prof. Madhura Vyawahare4,” Sentiment Analysis of Posts and Comments of OSN”, (IRJET).

[9] Lettura Exequiel Fuentes, Keith Norambuena Brian ,Claudio Meneses Villegas, ” Sentiment analysis and opinion mining applied to scientific paper reviews”,2019.

[10] Kajal , Prince Verma,” the sentimental analysis on product reviews of amazon data using the hybrid approach”, 2019 (IRJET).

[11] Dr. R. Ramachandiran , Arunnkumar , Balachander, ” Twitter Sentiment Analysis”, 2021(IRJET).

[12] Prof. S.J. Pachouly, GargeeRaut, KshamaBute, RushikeshTambe, ShrutiBhavsar, ”Depression Detection on Social Media Network (Twitter) using sentiment analysis”, 2021(IRJET).

[13] Arpit Upadhyay, Nishi Sharma, Aryan Chaudhary, Divya Jain, “Sentiment Analysis Of Generalized Text And Tweets”, 2020(IRJET).

[14] Princy Sharma1, Vibhakar Mansotra, "Social Media Sentiment Analysis: A Review", 2020 IRJET.

[15] Pratiksha Karpe, Abhishek Agarwal, Rakhi Marathe, Prachi Kolekar, Prof. Namrata Wasatkar, "Author Identification and Sentiment Analysis for novels using Natural Language Processing", 2020 IRJET