Machine Learning Technique Using Sentiment Analysis and Affective Computing Related Twitter Data.

1st Sumit Bansal
Dept. of Computer Science and Engineering
Gurukula Kangri (Deemed to be University) Haridwar, India

2nd Chirag Patel
Dept. of Computer Science and Engineering
Gurukula Kangri (Deemed to be University) Haridwar, India

3rd Ashish Bibyan
Dept. of Computer Science and Engineering
Gurukula Kangri (Deemed to be University) Haridwar, India

4th Kunal Singh Shekhawat
Dept. of Computer Science and Engineering
Gurukula Kangri (Deemed to be University) Haridwar, India

5th Divyank Singh
Dept. of Computer Science and Engineering
Gurukula Kangri (Deemed to be University) Haridwar, India

***Abstract*—The use of social media has led to a significant increase in the amount of user-generated data. With the growing availability of such data, there has been a surge in interest in using it for sentiment analysis and affective computing. In this paper, we propose a machine learning technique that utilizes sentiment analysis and affective computing to analyze Twitter data related to various topics. Our approach involves pre-processing the data, extracting relevant features, and then applying machine learning algorithms to classify the sentiment and emotions of the tweets. We demonstrate the effectiveness of our technique by analyzing Twitter data related to politics, sports, and entertainment. Our results show that our approach can accurately identify the sentiment and emotions expressed in tweets related to various topics. This technique can be applied in a variety of domains, such as marketing, politics, and public opinion analysis, to gain valuable insights into consumer attitudes and behavior.**

***Index Terms*—sentiment analysis, affective computing, machine learning algorithms, Twitter.**

I. Introduction

Social media platforms, such as Twitter, have become an increasingly popular means of communication, providing users with a platform to share their thoughts, opinions, and emotions. The vast amount of data generated by these platforms presents an opportunity for researchers and businesses to gain valuable insights into consumer behavior, public opinion, and sentiment analysis. Sentiment analysis is a technique that involves analyzing the emotional tone of a piece of text, such as a tweet, to determine whether the sentiment expressed is positive, negative, or neutral. Affective computing, on the other hand, involves the use of computational methods to recognize and interpret human emotions. Together, sentiment analysis and affective computing can be used to analyze social media data to gain a deeper understanding of user attitudes, preferences, and behavior.

In recent years, there has been growing interest in using machine learning techniques to perform sentiment analysis and affective computing on Twitter data. Machine learning algorithms can automatically learn patterns from data and use them to classify new data based on their features. This approach has been applied to various domains, including politics, marketing, and entertainment, to analyze sentiment and emotions expressed in tweets related to specific topics. However, the challenge lies in developing an effective machine learning model that can accurately classify tweets according to their sentiment and emotions.

In this paper, we propose a machine learning technique that uses sentiment analysis and affective computing to analyze Twitter data related to various topics. We present a step-by-step approach that involves pre-processing the data, extracting relevant features, and then applying machine learning algorithms to classify the sentiment and emotions of the tweets. We demonstrate the effectiveness of our approach by analyzing Twitter data related to politics, sports, and entertainment. The results show that our technique can accurately identify the sentiment and emotions expressed in tweets related to various topics. This technique can be applied in a variety of domains, such as marketing, politics, and public opinion analysis, to gain valuable insights into consumer attitudes and behavior.

Sentiment analysis has become an important field of study as it can provide insights into user preferences, opinions, and attitudes. It can help businesses in making data-driven decisions related to product development, marketing strategies, and customer service. It can also assist politicians in analyzing public opinion, and assist researchers in studying human behavior.

However, traditional methods of sentiment analysis have their limitations. They often rely on manual annotation or

keyword-based analysis, which may not capture the nuances of language and context. Furthermore, sentiment analysis alone may not be sufficient to gain a complete understanding of the user's emotions, as it may fail to capture emotions such as irony, sarcasm, or humor. Affective computing addresses these limitations by incorporating more advanced techniques, such as facial expression recognition and voice analysis, to recognize and interpret human emotions.

In recent years, there has been a growing interest in combining sentiment analysis and affective computing to analyze social media data. Twitter has emerged as a popular platform for sentiment analysis due to its real-time nature and vast amount of user-generated data. Machine learning algorithms have been used to develop models that can automatically classify tweets based on their sentiment and emotions. These models can be trained on labeled data, where the sentiment and emotions of the tweets are already known, and then applied to new, unlabeled data to classify them accordingly.

The proposed machine learning technique in this paper is based on a supervised learning approach, where the machine learning model is trained on labeled data. The data preprocessing involves removing stop words, handling negation, and converting the text into a numerical format. The relevant features are then extracted using techniques such as Bag of Words and TF-IDF. The machine learning algorithms used in the paper include Support Vector Machines, Decision Trees, and Random Forests.

The effectiveness of the proposed technique is demonstrated by analyzing Twitter data related to politics, sports, and entertainment. The results show that our approach can accurately classify the sentiment and emotions expressed in the tweets related to these topics. The proposed technique can be applied to a wide range of domains, such as marketing, politics, and public opinion analysis, to gain valuable insights into user behavior and attitudes.

II. METHODOLOGY

The methodology of our proposed machine learning technique involves a step-by-step approach that includes data preprocessing, feature extraction, and model training. We utilize a total of eight machine learning models to analyze the Twitter data and integrate them into an AWS server for efficient analysis.

To further explain our methodology, let us discuss each step in more detail.

1. Data Collection: We use the Twitter API to collect relevant tweets based on the topic of interest. The Twitter API provides a rich source of data for sentiment analysis as it allows us to collect data in real-time and in large volumes. We use the Tweepy library to connect to the Twitter API and collect tweets based on relevant keywords or hashtags.

2. Data Pre-processing: The collected data is pre-processed to remove irrelevant characters, such as emojis and special characters. We also convert the text to lowercase, remove stop words, and handle negation by adding the prefix 'not' to the words that follow a negation word. We also perform stemming or lemmatization to reduce the dimensionality of the data.

3. Feature Extraction: We use two feature extraction techniques, Bag of Words and TF-IDF, to extract relevant features from the pre-processed data. Bag of Words is a simple technique that represents text as a vector of word frequencies. TF-IDF is a more advanced technique that takes into account the frequency of a word in a document and the inverse frequency of the word in the entire corpus.

4. Data Splitting: We split the data into training and testing sets to evaluate the performance of the machine learning models. We use an 80-20 split, where 80

5. Model Selection: Our proposed methodology utilizes a total of eight machine learning models, including Support Vector Machines (SVM), Logistic Regression (LR), Bernoulli Naive Bayes (BNB), Light Gradient Boosting Machine (LGBM), Gradient Boosting Classifier (GBC), K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), and Random Forest Classifier (RFC). Our proposed methodology utilizes a total of eight machine learning models, including Support Vector Machines (SVM), Logistic Regression (LR), Bernoulli Naive Bayes (BNB), Light Gradient Boosting Machine (LGBM), Gradient Boosting Classifier (GBC), K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), and Random Forest Classifier (RFC).

Support Vector Machines (SVM): SVM is a popular machine learning model used for classification and regression analysis. It works by finding a hyperplane that maximizes the margin between the data points of different classes. SVM has been widely used in sentiment analysis due to its ability to handle high-dimensional data and non-linear boundaries. Logistic Regression (LR): LR is a popular machine learning model used for classification analysis. It works by estimating the probability of an event occurring using a logistic function. LR has been widely used in sentiment analysis due to its simplicity and interpretability. Bernoulli Naive Bayes (BNB): BNB is a variant of the Naive Bayes algorithm that assumes that the features are binary variables. It works by estimating the probabilities of each feature given the class and then multiplying them to obtain the probability of a particular class. BNB has been widely used in sentiment analysis due to its simplicity and efficiency. Light Gradient Boosting Machine (LGBM): LGBM is a gradient boosting algorithm that uses a tree-based model. It works by combining several weak learners to form a strong learner. LGBM has been widely used in sentiment analysis due to its ability to handle large volumes of data and its speed. Gradient Boosting Classifier (GBC): GBC is a variant of the gradient boosting algorithm that uses a decision tree-based model. It works by combining several weak decision trees to form a strong classifier. GBC has been widely used in sentiment analysis due to its high accuracy and ability to handle complex data. K-Nearest Neighbors (KNN): KNN is a popular machine learning model used for classification analysis. It works by finding the k-nearest neighbors of a data point and assigning it the class that occurs most frequently among its neighbors. KNN has been

widely used in sentiment analysis due to its simplicity and effectiveness. Stochastic Gradient Descent (SGD): SGD is a variant of the gradient descent algorithm that uses a randomly selected subset of the data for each iteration. It works by iteratively adjusting the weights of the model to minimize the loss function. SGD has been widely used in sentiment analysis due to its efficiency and scalability. Random Forest Classifier (RFC): RFC is a variant of the decision tree algorithm that uses a forest of decision trees. It works by combining several decision trees to form a strong classifier. RFC has been widely used in sentiment analysis due to its high accuracy and ability to handle complex data.

In our proposed methodology, we utilize these eight machine learning models to analyze Twitter data and classify the sentiment and emotions of the tweets. By using multiple models, we aim to improve the accuracy and robustness of the sentiment analysis. These models are selected based on their performance in previous studies and their ability to handle large volumes of data.

6. Model Training: We train the selected models on the preprocessed and feature extracted data using the training set. We use the scikit-learn library in Python for model training and evaluation.

7. Model Evaluation: We evaluate the performance of each model on the testing set using metrics such as accuracy, precision, recall, and F1-score. We also perform a comparative analysis of the performance of each model to select the best-performing model.

8. Model Integration on AWS Server: We integrate the eight machine learning models into an AWS server to enable efficient and real-time analysis of Twitter data. We use Amazon S3 for storing the data and Amazon EC2 for hosting the machine learning models. The integration of the models on AWS ensures fast and efficient analysis of large volumes of data.

In conclusion, our proposed methodology provides a robust approach for analyzing Twitter data related to various topics. The integration of the machine learning models in an AWS server ensures fast and efficient analysis of large volumes of data, which is essential for real-time analysis of Twitter data.