International Journal of Information Management Data Insights 1 (2021) 100019

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

# Sentiment analysis and classification of Indian farmers' protest using twitter data

Ashwin Sanjay Neogi[a], Kirti Anilkumar Garg[a], Ram Krishn Mishra[a,*], Yogesh K Dwivedi[b]

[a] Department of Computer Science, BITS Pilani, Dubai Campus, Dubai, United Arab Emirates
[b] Emerging Markets Research Centre (EMaRC), School of Management, Swansea University Bay Campus, Swansea, SA1 8EN, Wales, UK

ARTICLE INFO

*Keywords:*
Sentiment analysis
TF-IDF
Bag-of-words
Machine learning
Farmers' protest

ABSTRACT

Protests are an integral part of democracy and an important source for citizens to convey their demands and/or dissatisfaction to the government. As citizens become more aware of their rights, there has been an increasing number of protests all over the world for various reasons. With the advancement of technology, there has also been an exponential rise in the use of social media to exchange information and ideas. In this research, we gathered data from the microblogging website Twitter concerning farmers' protest to understand the sentiments that the public shared on an international level. We used models to categorize and analyze the sentiments based on a collection of around 20,000 tweets on the protest. We conducted our analysis using Bag of Words and TF-IDF and discovered that Bag of Words performed better than TF-IDF. In addition, we also used Naive Bayes, Decision Trees, Random Forests, and Support Vector Machines and also discovered that Random Forest had the highest classification accuracy.

## 1. Introduction

The farmers' protest in India is ongoing in the northern parts of India against the three farm acts passed by Parliament in September 2020. The three acts are: The Farmers' Produce Trade and Commerce (Promotion and Facilitation) Act, The Farmers' (Empowerment and Protection) Agreement of Price Assurance and Farm Services Act, and The Essential Commodities (Amendment) Act. These Acts, according to the government, will "change Indian agriculture" and "attract private investment." The Farmers' (Empowerment and Protection) Agreement on Price Assurance and Farm Services Act, 2020, establishes contract farming, in which farmers produce crops in exchange for a mutually negotiated remuneration under contracts with corporate investors.Protesting farmers are concerned that powerful investors will bind them to unfavorable contracts created by major corporate law firms, with liability clauses that, in most circumstances, are beyond the comprehension of poor farmers.

More than 40,000 protesters have committed themselves to ensure that the three acts are recalled as per their demand. The farmers have rejected the government proposal of suspending the laws for 18 months and the government has insisted that the protests are a result of misinformation.

The protests are an integral part of a democratic society and they can be fundamental in shaping the future of the society. When one community protests, the entire society comes forwards in support of the community or to voice their own opinions against the community. This plays a crucial role in society's development. Many of such protests have been instrumental in discarding age-old beliefs that were not relevant in the current society. The protests also enable common people to be heard by their elected leader. At the same time, some form of protests can also cause violence and create an imbalance in society. It is important to understand the emotions behind online conversations to understand a protest because this allows us to take into consideration a broader audience and be inclusive of both direct and indirect participants.

As a result of the widespread protest, there has been an influx of opinions and emotions shared by people on social media on an international level. The opinions and sentiments of the public were across a diverse range. The farmers received support from all over the world with thousands of people expressing their opinions on social media. Hashtags like #farmersprotest, #iamwithfarmers, #SpeakUpForFarmers, #IStandWithFarmers and #kisanektazindabaad were trending on twitter. However, lots of groups came forward and called it anti-national propaganda as well. Some groups also believed that the farm bills are in the favour of the farmers and the protests are being held because of a lack of information.

In this fast-pacing world, information spreads digitally between users and it can also shape the way other users feel about an event. Therefore, it is crucial to understand the sentiment of the masses. Sentiment analysis is a technique of processing natural language to analyze and under-

* Corresponding author.
*E-mail address:* rkmishra@dubai.bits-pilani.ac.in (R.K. Mishra).

https://doi.org/10.1016/j.jjimei.2021.100019
Received 15 May 2021; Received in revised form 4 June 2021; Accepted 4 June 2021

stand human emotions (Pak & Paroubek, 2010; Pietra, Berger, & Pietra, 1996). We have used sentiment analysis to analyze twitter textual data and it gives us an understanding of two important metrics: polarity and subjectivity (Srivastava, Singh, & Drall, 2019). The polarity of the data ranges from -1 to +1, with -1 being a completely negative emotion, 0 being neutral and +1 being completely positive emotion. The subjectivity of data ranges from 0 to 1, with 0 being a complete opinion and 1 being a fact (Pang & Lee, 2004).

With this research, we aim to analyze and understand the sentiment of the masses regarding farmers' protest. Bugden (2020) We extracted 150 tweets related to the protest from each day starting from 4 No. 2020 to 5 Mar 2021, using the hashtag keyword 'farmers protest'. The main objective of this research is to understand the sentiments of the public on farmers' protest shared on the microblogging website Twitter. Go, Huang, and Bhayani (2009) Our objective is to understand the sentiments of the Indian citizens towards the three acts passed by the government by incorporating NLP techniques. In addition, we also aim to analyze the polarity and factuality of Twitter data regarding the demonstrations by extracting twitter data. We plan to use visualization libraries to conduct a thorough analysis of Twitter data. We will also determine the study's obstacles and problems and discuss the contribution of this research in possible future works.

We have used Bag of Words and TF-IDF to convert the textual information to numeric weightage in vector format. Furthermore, we used four classifiers namely Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine for prediction purposes.

In the next section, we discuss some of the previous works related to sentiment analysis of Twitter data. In Sections 3 and 4, we discuss methodology and Model Building. We discuss the Model Prediction in Section 5 and experimental results and analysis in Section 6. We have mentioned Discussion in Section 7. The conclusion and the future scope of the research is covered towards the end in Section 8.

## 2. Related work

Twitter is a hub where substantial amounts of data are being generated by users. It is reported that Twitter users generate 12 GB of data every day. It is widely utilized by the general people, who utilize it to express their opinions on a variety of public topics and to voice their complaints to businesses and government agencies. Twitter, being a social networking platform, generates data which may be used for a number of things, including analysis on certain subjects, people, and so on. With the rise of AI, we now have state-of-the-art machine learning and natural language processing algorithms at our disposal, which can be used to analyze the emotions of users' conversations on social media platforms, where it has become a key instrument in understanding human behavior, studying public relations, and helping to solve various issues (Chintalapudi, Battineni, Canio, Sagaro, & Amenta, 2021).

We can clearly see that sentiment analysis is getting more and more popular as e-commerce, SaaS solutions, and digital technologies are advancing everyday. There are numerous applications to sentiment analysis and various researchers have used sentiment analysis to derive explanations from the Twitter data to examine an event or to solve an issue. There are some intriguing projects underway in this area, such as (Sarlan, Nadam, & Basri, 2014) Zainab Tariq Soomro et al. analyzed over 18 million tweets related to novel coronavirus. In Soomro, Ilyas, and Yaqub (2020). The tweets were investigated to see if there was a link between public mood and the number of coronavirus infections rising or falling. Users make comments in a variety of languages as the Internet spreads throughout the world. Sentiment analysis in a single language raises the danger of missing important information in other languages' texts. Multilingual sentiment analysis approaches have been created to analyze data in many languages. Abdul-Mageed and Diab (2014) used sentiment analysis to understand sentiments in the Arabic language largely using Twitter tweets and Youtube comments. Similarly, in Garcia and Berton (2021), Klaifer Garcia et al. uses sentiment analysis in English and Portuguese to understand the impact of the pandemic in two geographical locations namely Brazil and the USA using Twitter data. Another application is in Mihalcea, Banea, and Wiebe (2010) Efthymios Kouloumpis et al. uses sentiment analysis in informal languages to understand and evaluate the usefulness of sentiment analysis in Kouloumpis, Wilson, and Moore (2011). Lastly, Ram Krishn Mishra et al. has analyzed sentiments of various reviews to create a hotel recommendation system in Mishra, Urolagin, and Jothi (2019). In a nutshell, Sentiment analysis may be utilized for a variety of purposes, ranging from service recommendations to problem solutions.

Sentiment analysis can also have multiple layers embedded into it. Many researchers have implemented multiple dimensions to sentiment analysis to improve and upgrade the technique. T. Wilson et al. has worked on phrase-level sentiment analysis to determine sentiments behind contradictory statements in Jain and Nemade (2010). In Pang, Lee, and Vaithyanathan (2002), B. Pang et al. use sentiment analysis to identify the exact positive (thumbs-up) or negative (thumbs-down) sentiment behind a text. A. Pak et al. has researched to automatically find a corpus and use it for sentiment analysis in Pak and Paroubek (2010).

B. Liu et al. have surveyed in Zhang, Ghosh, Dekhil, Hsu, and Liu (2011) where they highlight how humans fail to unbiasedly assess a sentiment as they tend to pay attention to their preferences. Cambria, Olsher, and Rajagopal (2014) They then conduct a survey analysis of opinion mining and sentiment analysis. T. Nasukawa also researches sentiment analysis of specific sections of the document in Nasukawa and Yi (2003), rather than categorizing the entire document as one sentiment. R.K. Bakshi et al. has done an interesting study on sentiment analysis to understand how sentiments expressed in a tweet affect the stock price of a multinational firm in Bakshi, Kaur, Kaur, and Kaur (2016). In Maas et al. (2011), A.L. Maas et al. use both supervised and unsupervised learning to learn word vectors for sentiment analysis. R. Prabowo uses hybrid classification in Prabowo and Thelwall (2009) to improve the accuracy of sentiment analysis.

Furthermore, these social networking websites always have a breeding ground for misinformation, propaganda and fake news. Over the past decade, the use of political Twitter accounts has skyrocketed. Political leaders play a vital role in influencing the common public about the goals and issues (Grover, Kar, Gupta, & Modgil, 2021). Due to this, There might arise problems of misinformation, in Aswani, Kar, and Ilavarasan (2019), the authors have presented their findings indicating that the tweet emotion and polarity plays a significant role in determining whether the shared content is authentic or not. Therefore, there should be a proper governance policy and management of big data where users are responsible for generating content (Sarin, Kar, & Ilavarasan, 2021) as it can easily be misused and propagated in a rapid manner as mentioned in Joseph, Kar, and Ilavarasan (2021).

## 3. Methodology

Here we discuss the detailed approach that we have followed to discover the sentiments of people about the ongoing farmers' protest in India. Fig. 1 shows the progressive steps we have taken to analyze and predict the sentiment of a particular Twitter user. The process starts with collecting the data from Twitter followed by a few crucial components such as cleaning and preprocessing the data to bring it to a machine-understandable format (Kotsiantis & Kanellopoulos, 2006). Further we move on to calculate and classify the sentiment of a user based on two parameters. Also, visualization methods can be used to analyze the sentiments on various factors. Lastly, we use machine learning algorithms to predict the tweets and plot the performance metrics.

### 3.1. Dataset

A total of 18,000 tweets have been collected over a period of four months. The raw data has been collected by us using an open-source python library called tweepy to directly access the Twitter API which

**Table 1**
Sample Dataset of Farmers Protest .

| Datetime | Tweet Id | Tweets | Username |
|---|---|---|---|
| 2020-11-05 14:51:51 | 1324360000000000000 | Farmers shout slogans as they block road during the protest against farms bills by center government at the entrance gate in Amritsar PHOTOS-PRABHJOT GILL https://t.co/SOY9WL | FW_Delhi_Chd |
| 2020-12-21 23:53:42 | 1341170000000000000 | Farmers from the nearby Indian states of Punjab, Haryana and Uttar Pradesh are organizing on a national level to protest the agricultural policies set forth by the Bharatiya Janata Party. #IrisNews #ThelrisNews #NewDelhi #Farmers #Agriculture #Protest https://t.co/kDOHCDA8aC | ThelrisNYC |
| 2021-01-14 23:01:04 | 1349850000000000000 | Rahul Gandhi is inciting the farmers protest. Public life around the Singhu border is getting impatient as they can't open their shops and are even unable to come out of their houses. Their losses should be recovered from Rahul Gandhi. @HMOlndia @PMOlndia https://t.co/ypREvSS7It | HarashKhatana |
| 2021-02-19 23:03:05 | 1362900000000000000 | Absolutely disgusting and shocking. Please raise your voices against this fascism #FarmersProtest #ReleaseDetainedFarmers https://t.co/mKGNzjKD69 | NijjarKash |

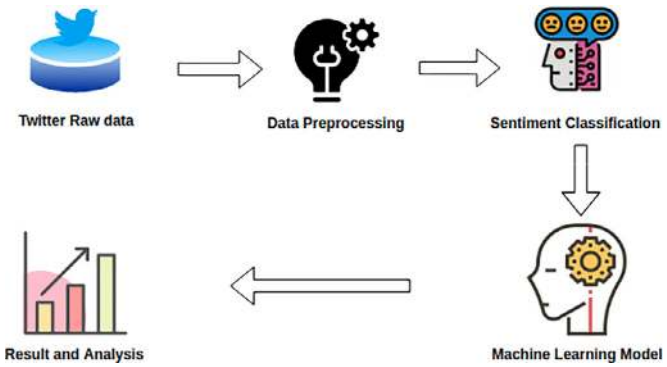


**Fig. 1.** Sentiment Analysis Step by Step Approach.

in turn uses private access tokens and consumer keys for authentication purposes. Since the farmers' protest began around the month of November 2020, we chose the starting date as 5th November and the ending date as 5th March 2021. The DateTime library was incorporated for a customized script which was written to explicitly retrieve 150 tweets per day and used to store them in a python list. We used the keyword 'farmers protest' as a search query wherein all the tweets which contained the words "farmers", "protest", and "farmers protest" were conglomerated together (Kumar, Kar, & Ilavarasan, 2021). The tweets were gathered from 9061 unique users and stored in a CSV format. The dataset consists of 4 attributes namely Datetime, Tweet Id, Tweets, and the Username and a sample is shown in Table 1.

### 3.2. Data preprocessing

One of the most important aspects of analyzing data is to ensure that our data is being understood by machines. Machines do not understand text, images, or videos, they can comprehend only 1's and 0's. To be able to provide an input consisting of 1's and 0's is a multistep process. Pre-processing the data is an absolute necessity and calls for a technique called data cleaning which involves transforming raw data into a machine-understandable format (Nithya, 2016). Since we possess a huge text dataset consisting of tweets, we need to clean it to remove certain discrepancies to avoid inconsistent data. Our approach to cleaning data is fairly simple. Before we could start with data cleaning, we took advantage of excel's in-built option of removing duplicate records present in our data. We noticed that there were 835 duplicate entries and as a result, the shape of the dataset or the number of tweets reduced to 17,165 unique tweets. We started by removing @-mentions, Retweets represented as 'RT', links, and hashtags symbols using regular expressions as these things do not add any sort of value. A point to be noted here is that we made sure to not remove any words after the hashtag as it can contain a valuable reference to the sentiment of the tweet. For example: in #istandwithfarmer, even though the symbol '#' does not add any positive or negative value to our analysis, the text "I stand with farmers" gives us insight into the state of the mind of the user. Further, the tweets containing special characters, punctuation, numbers, and emoticons were also removed.

Tokenization is defined as separating huge quantities of text into smaller units called tokens (Webster & Kit, 1992). Tokenization is a fundamental step in modeling text data. It helps in understanding the meaning behind the text by analyzing the sequence of the words. We used the porter stemmer to reduce the inflection towards their root forms. This was done by stripping the suffix to produce stems (Jivani, 2011). Lastly, the fully pre-processed tweets were stored in a new pandas column called "Cleaned_Tweets" in our existing data frame of tweets dataset.

### 3.3. Lexicon based sentiment calculation

In our project, to avoid the process of generating labeled data we have decided to apply a lexicon-based approach. This approach requires calculating the semantic orientation of words present in the text (Rajman & Besancon, 1998). The main advantage of choosing a lexicon-based approach is that it is much simpler to understand and can easily be modified by a human. Using this technique, the semantic orientation can be captured and labeled as neutral, positive, or negative. From a definition perspective, sentiment analysis is a method to retrieve subjectivity and polarity from text and on the other hand, semantic orientation measures the polarity and strength of the text (Szabolcsi, 2004). In this line of approach, adjectives and adverbs are used to disclose the semantic orientation of the text (Jain, Seeja, & Jindal, 2021) (in our case it is a tweet). The next step is to calculate the sentiment orientation value considering the combinations of adverbs and adjectives. Furthermore, a single source for the whole value is at hand. A popular Python package called TextBlob is applied which supports complex operations and analysis on the tweet data. The tweets are represented by a numerical format and TextBlob assigns individual scores to all the tweets. Lastly, the sentiment of the tweets is calculated by a pooling operation wherein it takes the average of all the sentiments.

### 3.4. Sentiment classification

We also need to understand that TextBlob returns two values and those are polarity and subjectivity of the tweet. Polarity lies between [−1,1], −1 defines a negative sentiment and 1 defines a positive sentiment. Negation words reverse polarity causing it to fall below 0. Subjectivity lies between [0,1]. Subjectivity tells us the amount of personal opinion and factual information present in the tweet. When the subjectivity is high it indicates there is more personal opinion. TextBlob has one more parameter and that is intensity. TextBlob calculates subjectivity by looking at the intensity. Intensity determines if a word has any sort of influence on the next word. For English, adverbs are used as modifiers such as 'very good' which is explained in the previous section that it uses a lexicon-based approach. There are cases where the values are

**Table 2**
Top positive and negative tweets from TextBlob.

| Tweet | Polarity Score | Sentiment |
|---|---|---|
| if it is so anti farmer why are the protests isolated to very few states like AP, TN have highest no of farmers and not a single protest, none of them r bjp ruled even in Punjab,haryana there are farmers who appreciated the bill when asked by anti bjp media like ndtv | -0.0081 | Negative |
| Railways have suffered loss of Rs 2100 crores due to farmer protest in Punjab: | -0.0625 | Negative |
| Farmers commit suicide due to water scarcity and people like you are directly responsible for farmer deaths in Maharashtra. | -0.0166 | Negative |
| We have been serving farmers with their essential needs since they started protesting againsl farm laws in India. thanks to our heroic volunteers. | 0.3 | Positive |
| #FarmersProtest #StandWithFarmers STAY UNITED STAY PEACEFUL | 0.25 | Positive |
| The true measure of the JUSTICE of a system is the amount of protection it guarantees to the weakest. Our FARMERS are suffering Speak up for FARMERS! #FarmersProtest | 0.4375 | Positive |

exactly equal to 0. Based on the polarity value, a sentiment score will be assigned, and the computation is calculated in such a way that if the score is less than 0, the sentiment is returned as negative. If the polarity is greater than 0, then the sentiment is returned as positive. In all other cases, the score will be 0 and the sentiment is returned as neutral. After classifying the sentiments, we took a look at the count of the number of sentiments and found that a large number of people (8253) have neutral feelings about the protest indicating that neither they support the farmers' protest nor they support the government (Table 2).

## 4. Model building

This section covers the sentiment classification and prediction of tweets using 4 popular supervised machine learning algorithms viz. Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine. Since computers cannot process text data in its raw form, we must prepare the data before training the machine learning models. The text must be manually decomposed into a numerical format that the computer can understand (Rawat, Rawat, Kumar, & Sabitha, 2021). Hence, we examine the results obtained using two natural language processing techniques such as Bag of Words and Term Frequency and Inverse document frequency approach (Zhang, Jin, & Zhou, 2010). Both BoW and TF-IDF are NLP techniques that help us convert tweet sentences into numeric vectors (Manning et al., 2014).

### 4.1. Bag of words and term frequency - inverse document frequency

#### 4.1.1. Bag of words

The Bag of Words model is a technique of extracting features from a text that can be used in modeling, like in our scenario for machine learning algorithms for tweet sentiment classification. In simplistic terms, it is a group of words used to describe a sentence in a text with word count. It involves two things: first a list of well-known terms, and second metric for determining the presence of well-known terms. Another thing about BoW is the order in which they appear is discarded.

The first step is to construct a vocabulary out of all the distinct words in our tweets Data frame. The next step is to list each of these distinct words and monitor their occurrence in every single tweet. Finally, you pass the matrix of numbers to the model for training purposes.

#### 4.1.2. TF-IDF

The TF-IDF system outperforms the BoW approach because it is used to evaluate the importance of a word in a tweet (Aizawa, 2003). When scoring word frequency, a common issue is that highly recurrent terms begin to dominate the text, but it may lack the 'informational content' required for the model to correctly differentiate. The IDF is a metric for measuring the significance of a word. We need the IDF value since just computing the TF isn't enough to appreciate the significance of words: The Eq. (1) shows the calculation of term frequency of the term t in document d. Term Frequency is a score dependent on the frequency in which a term appears in the document. Inverse Document Frequency is a metric for determining how rare a word is based on a document.

$$TF(t,d) = \frac{N(t,d)}{T} \tag{1}$$

Here, TF(t,d) represents the term frequency of the term t in document d, N(t,d) is the number of times the term t appears in the document d, and T is the total number of terms in the document.

Thus, for each document and word, a different TF(t,d) value will be assigned.

$$IDF(t) = logN/(N(t)) \tag{2}$$

Equation (2) shows the calculation of IDF(t), which is the inverse document frequency of term t, N is the no, of documents, N(t) is the no. of documents with the term t.

$$TF - IDF = TF * IDF \tag{3}$$

Equation (3) gives the calculation of TF-IDF.

### 4.2. Naives bayes

The supervised learning algorithm, Naive Bayes is based on the Bayes' Theorem, which implies predictor independence (Yang, 2018). In simple terms, a Naive Bayes classifier assumes that the presence of one function in a class has no bearing on the presence of any other feature. This allows one to comprehend what the Bayes theorem says. Often in machine learning, we need to select the best hypothesis(h) given the dataset (d). One of the simplest ways to choose a hypothesis is to use our previous knowledge of the situation. The Bayes' Theorem allows one to quantify the likelihood of a hypothesis given prior knowledge (Zervoudakis, Marakakis, Kondylakis, & Goumas, 2021). Bayes' Theorem is stated as:

$$P(h/d) = (P(d/h) * P(h))/(P(d)) \tag{4}$$

Equation (4) gives the value for P(h/d), which is the probability of hypothesis h given the data d. This is called the posterior probability. P(d/h) is the probability of data d given that hypothesis h was true. P(h) is the probability of hypothesis h being true. This is called the prior probability of h. P(d) is the probability of the data.

We should choose the hypothesis with the highest probability after determining the posterior probability for each hypothesis. The Maximum Posteriori Probability (MAP) hypothesis is used to describe this. We used the sci-kit-learn library to implement the Naive Bayes algorithm and before that, we converted the tweets to a matrix of token counts using a count vectorizer.

### 4.3. Decision tree

In a tree-structured classifier, decision trees consist of internal nodes, which represent dataset attributes, branches that represent decision rules, and leaf nodes representing the outcome. In a decision tree, we have the decision node and the leaf node. Decision nodes are used to make decisions and have several branches, while leaf nodes are the result of those decisions and tell us whether the sentiment is positive, negative, or neutral and have no additional branches. Initially, our dataset

**Table 3**
Accuracy Chart of ML algorithms achieved through sentiment analysis.

| | Accuracy | |
|---|---|---|
| | Bag of Words | TF-IDF |
| **Naïve Bayes** | 72.9 | 71.33 |
| **Decision Tree** | 79.78 | 77.62 |
| **Random Forest** | 96.62 | 95.51 |
| **SVC** | 83.45 | 83.04 |

consisting of tweets is considered as the root node or as the starting point to gain information.

Entropy, which governs how a Decision Tree chooses to divide the results, is used as an algorithm in Decision Trees (Swain & Hauska, 1977). It has an effect on how a Decision Tree draws its boundaries. It's also worth noting that entropy values vary from 0 to 1 (Myles, Feudale, Liu, Woody, & Brown, 2004). Equation (5) explains the calculation of Entropy.

$$H(s) = -probabilityof log_2(p) - (-probabilityof log_2(n)) \quad (5)$$

where, p implies percentage of positive class and n implies percentage of negative class.

### 4.4. Random forest

Random forest is a supervised machine learning algorithm as well. A random forest is simply a set of decision trees. The Random Forest algorithm has two stages: random forest construction and prediction using the random forest classifier generated in the first stage (Biau & Scornet, 2016). a. Select "K" features at random from a total of "m" features such that k<<m. b. Calculate the node "d" using the best split point of the "K" functions. c. Using the optimal break, divide the network into daughter nodes d. Repeat measures 1 to 3 until the l number of nodes is reached. e. Build a forest by repeating steps 1 to 5 for "n" number of times to create an "n" number of trees.

The reason we used random forest was to see how much better the precision would be relative to the decision tree algorithm. With about 150–200 estimators, we discovered that the precision improved by about 23.07 percent as compared to the decision tree.

### 4.5. Support vector machine

Given a set of training examples that are each labeled as belonging to one of two categories, an SVM training algorithm generates a model that assigns new examples to one of two categories, making it a non-probabilistic conditional linear classifier. The aim of using SVMs is to find the best line in two dimensions or the best hyperplane in more than two dimensions to assist us in classifying our space. In both types of data, the maximum margin, or the maximum distance between data points, is used to locate the hyperplane (line).

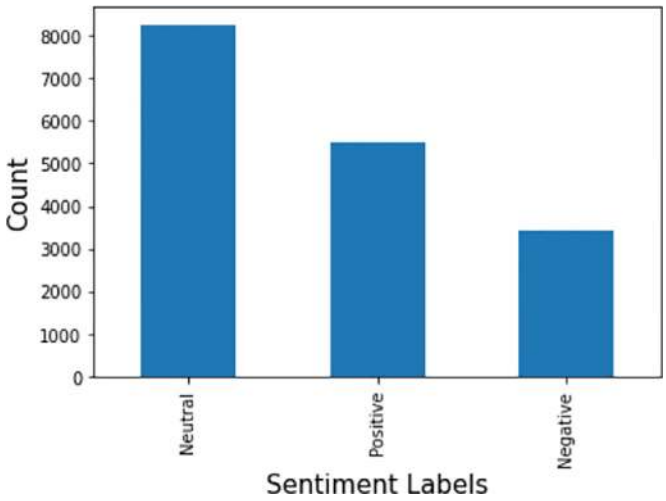


**Fig. 2.** Bar plot of Sentiment Counts.

## 5. Model prediction

Here we use a python-based machine learning library called sci-kit-learn to fit the above-mentioned algorithms into our dataset. This section also covers the prediction, visualization, and analysis of the results gathered.

Table 3 represents a comparison between accuracies of BOW and TF-IDF precision. There is just a minor disparity between the two, as we can see.

The Table 4 represents the precision, recall and f1-score for each class i.e., negative, neutral, and positive, belonging to each algorithm i.e., Naive Bayes, Decision Tree, Random Forest and SVC using Bag of Words.

The Table 5 represents the precision, recall and f1-score for each class of different algorithms using TF-IDF vectorizer.

Figure 3 depicts the flowchart of the entire sentiment analysis model procedure. The tweets dataset is first sent through data pre-processing, which involves cleaning the data by removing stop words and dupli-

**Table 4**
Precision, Recall and F1 score in each class for all ML algorithms using Bag of Words.

| | Negative | | | Neutral | | | Positive | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | fl-score | Precision | Recall | fl-score | Precision | Recall | fl-score |
| **Naïve Bayes** | 0.55 | 0.65 | 0.59 | 0.8 | 0.78 | 0.79 | 0.73 | 0.68 | 0.7 |
| **Decision Tree** | 0.73 | 0.65 | 0.69 | 0.83 | 0.89 | 0.86 | 0.78 | 0.76 | 0.77 |
| **Random Forest** | 0.99 | 0.92 | 0.95 | 0.95 | 0.99 | 0.97 | 0.98 | 0.95 | 0.96 |
| **SVC** | 0.83 | 0.65 | 0.73 | 0.82 | 0.98 | 0.89 | 0.87 | 0.73 | 0.79 |

**Table 5**
Precision, Recall and F1 score in each class for all ML algorithms using TF-IDF vectorizer

| | Negative | | | Neutral | | | Positive | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | fl-score | Precision | Recall | fl-score | Precision | Recall | fl-score |
| **Naïve Bayes** | 0.88 | 0.31 | 0.46 | 0.71 | 0.92 | 0.8 | 0.72 | 0.71 | 0.71 |
| **Decision Tree** | 0.71 | 0.63 | 0.67 | 0.83 | 0.88 | 0.86 | 0.76 | 0.74 | 0.75 |
| **Random Forest** | 0.96 | 0.93 | 0.94 | 0.97 | 0.98 | 0.97 | 0.95 | 0.96 | 0.96 |
| **SVC** | 0.84 | 0.6 | 0.7 | 0.8 | 0.97 | 0.88 | 0.89 | 0.76 | 0.82 |

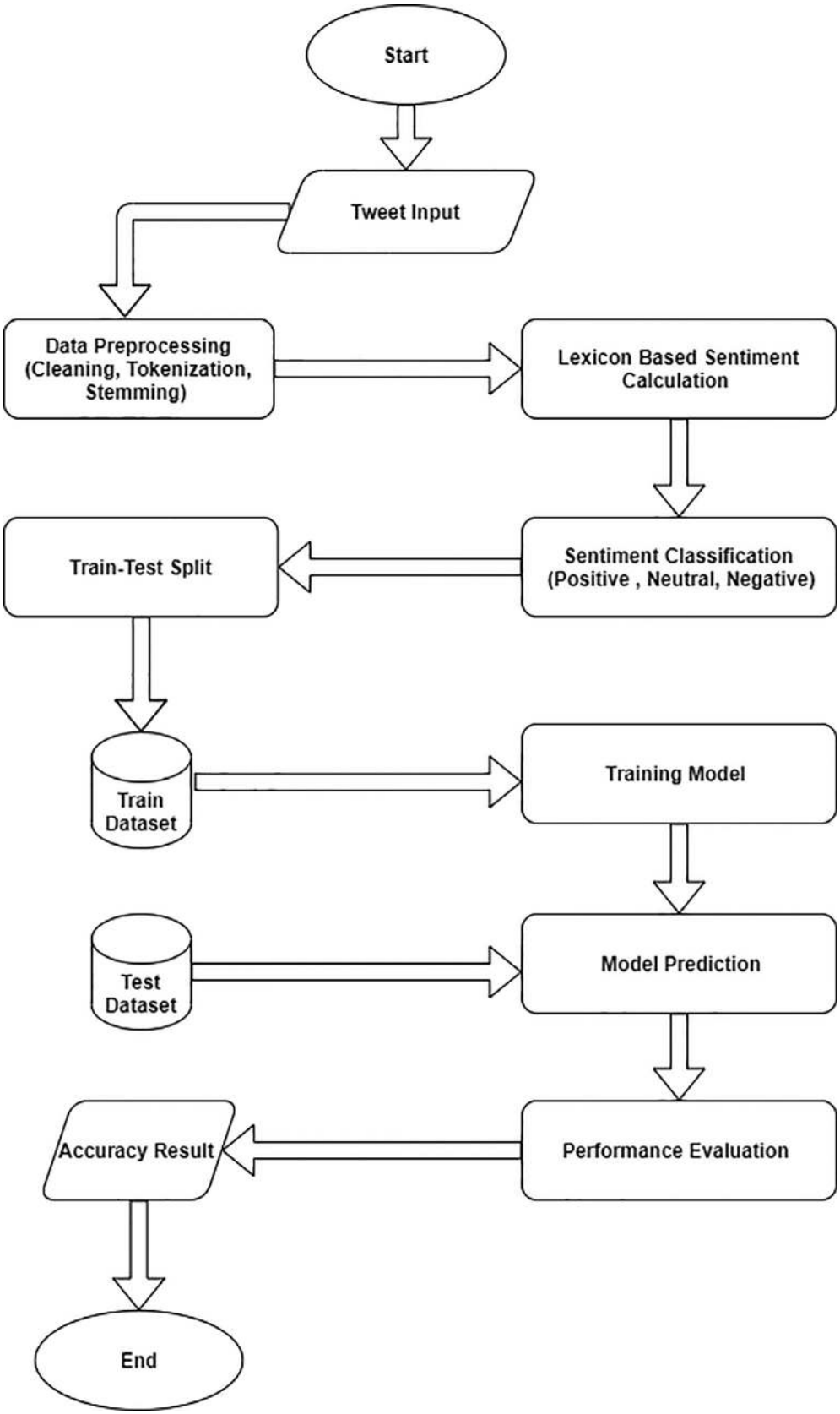


**Fig. 3.** Scatter Plot of Polarity and Subjectivity.

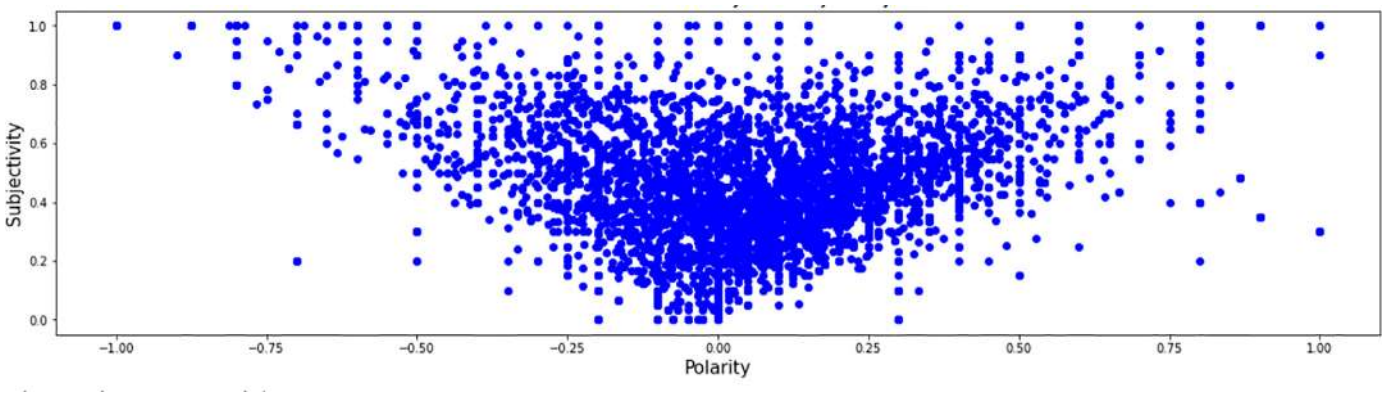


**Fig. 4.** Scatter Plot of Polarity and Subjectivity.

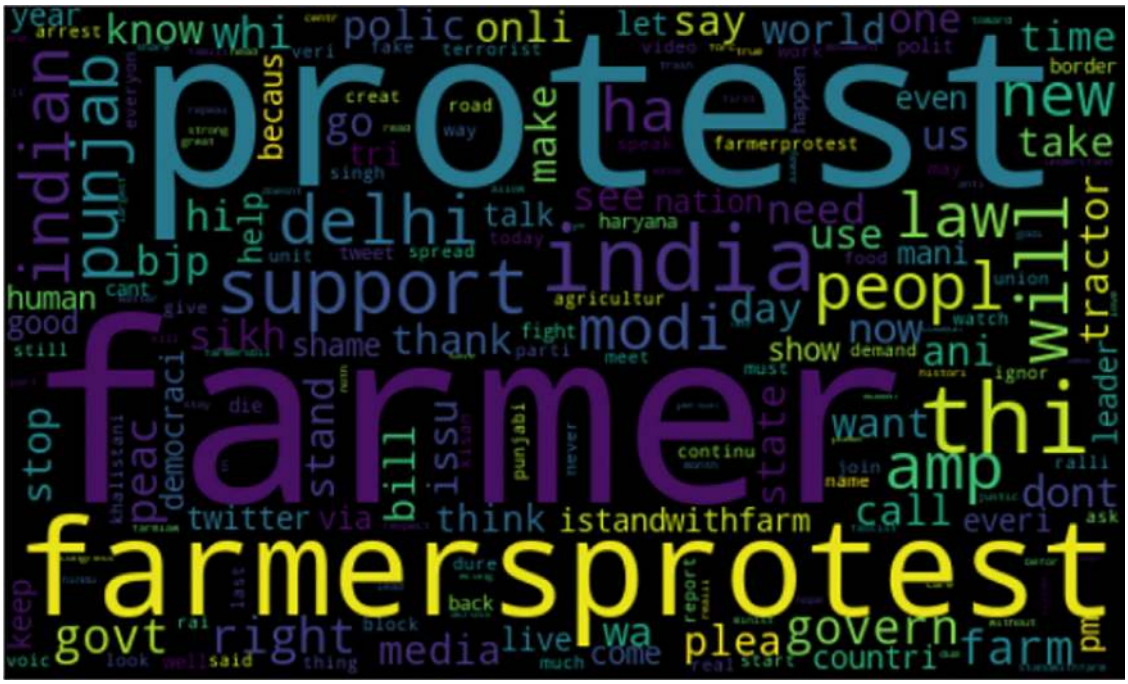


**Fig. 5.** Word Cloud.

cation, then tokenizing the dataset, and stemming the word to its root form. Following pre-processing, the data is fed into a lexicon-based sentiment calculation model, which aids in determining the text's semantic orientation. TextBlob is used to assign polarity and subjectivity values to tweets. The sentiments are then manually classified depending on the polarity value. We then use two NLP approaches called Bag of Words and TF-IDF to translate the tweets into a numerical format. We split the dataset into training and testing after converting it to a numeric format, and we feed the training dataset to machine learning models. We use the test dataset to predict attitudes once the model has learned from the data. Finally, we assess the models using performance indicators as precision, recall, accuracy, and confusion matrix.

## 6. Experimental results and analysis

Figure 4 represents a plot of polarity and subjectivity for all the tweets. While the polarity is concentrated mostly in the center, the subjectivity is spread out across the graph. This indicates that our collection of tweets shows a wide range of subjectivity and most of the tweets fall in [-0.75,0.75] polarity scale implying that the extremely negative or positive sentiments are significantly low. While the users have shared their complete opinions as well as facts about farmers' protest online, most of the tweets show a mid-range of negative and positive sentiments. Extreme language has been used by an exceptionally few users while sharing their sentiments of farmers' protest. In the graph, tweets with low subjectivity are concentrated at the center of the polarity range [-1, +1] and the tweets with high subjectivity are scattered across the polarity range [-1, +1]. This implies that the facts in the farmers' protest were neutral, however, the sentiments expressed by the users were across a varied range of negative to positive. This is understandable because a fact (low subjectivity) is more likely to be neutral (with polarity 0) and an opinion (high subjectivity) is more likely to have a diverse range of negative to positive sentiments.

Figure 5 shows the word cloud obtained from the cleaned tweets. Word Cloud is a pictorial representation of commonly used words in a particular dataset. We provided our dataset of cleaned tweets to the model to generate this word cloud (Heimerl, Lohmann, Lange, & Ertl, 2014). The entire word cloud represents the most frequently used words. The words with a larger font occur more commonly than the words with a smaller font. The word cloud can give us an overview of the farmers'

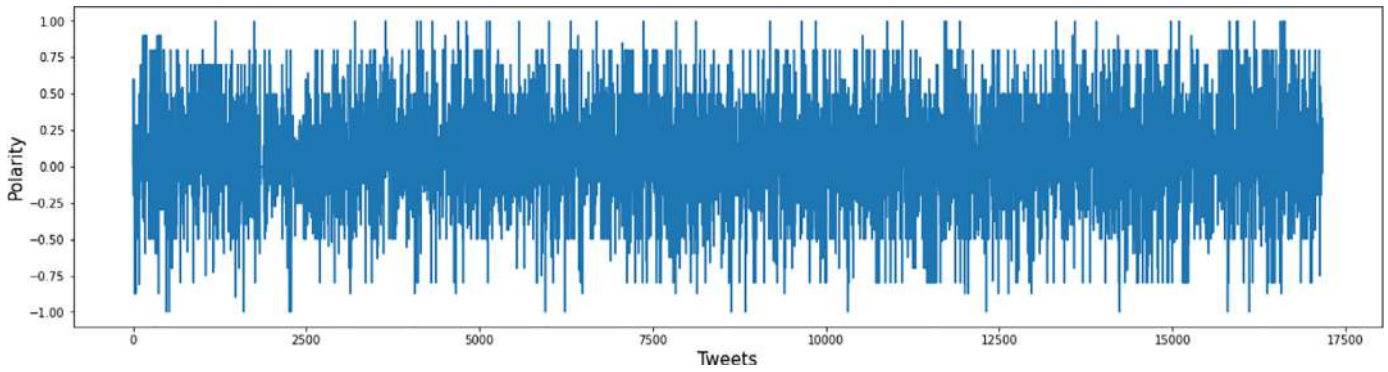


**Fig. 6.** Polarity scale of each tweet in the dataset.

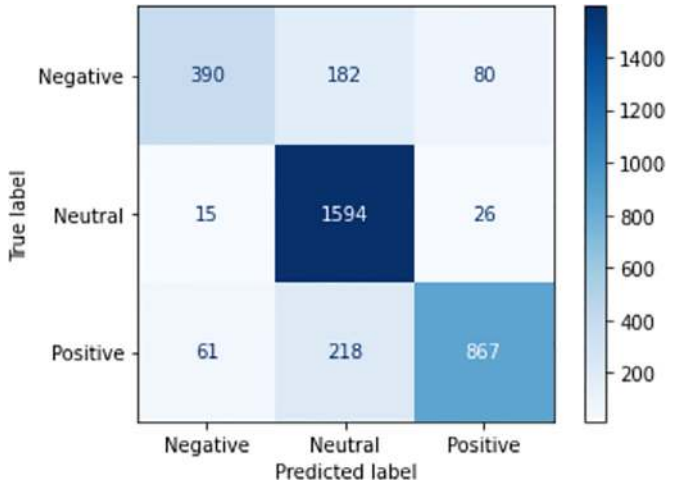


**Fig. 7.** Confusion Matrix for Random Forest using TF-IDF Vectorizer.

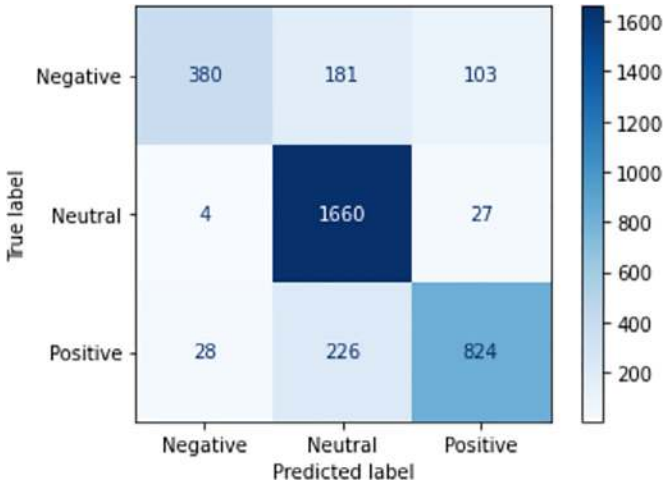


**Fig. 8.** Confusion Matrix for SVC using TF-IDF vectorizer.

protest. It can also help us in understanding the essence of the protest. In our word cloud of cleaned tweets, the words that occur most frequently are farmer, protest, support, India, Modi, Delhi, Indian, Punjab and many others. While the words like farmer and protest bring to light the common motive of the tweets, the words like support, Istandwithfarmer indicate that people were tweeting in the support of the farmers. The words india, Modi, Delhi, Indian and Punjab show that the protest was centered in India in the region of Delhi and Punjab. A copious number of words like help, plea, govern, peace and democracy have been mentioned which reveal that Twitter users expected the governments interference to meet the demands of the farmers.

Figure 7 depicts the confusion matrix for Random Forest using TF-IDF Vectorizer. The true label mentions the actual sentiment of the tweet and the predicted label is the predicted sentiment of the tweet. The confusion matrix implies that 380 + 1660 + 824=2864 tweets were predicted with the same sentiment as their true label. It also implies that 181 + 103 + 27 + 4 + 28 + 226 = 569 tweets were predicted with a wrong sentiment. 83% of the tweets were predicted correctly as their true label while the rest were predicted with a wrong sentiment compared to its true label.

Similarly, Fig. 8 depicts the confusion matrix for SVC using TF-IDF vectorizer. According to the confusion matrix, 390 + 1594 + 867 = 2851 tweets were predicted correctly while 582 tweets were predicted incorrectly.

Figure 9 shows the accuracy of various ML algorithms during sentiment analysis of farmers' protest tweets (Huang & Ling, 2005). Naive Bayes showed the minimum accuracy at 72%, while Random Forest has the highest accuracy of 96.6%. Random Forest received the highest accuracy because it is an ensemble of decision tree algorithms. Decision Tree and SVM fall in the middle with an accuracy of 79.8% and 83.5% respectively.

Figures 10 and 11 show the histograms of negative and positive subjectivity correspondingly. Here the x-axis is the distribution of the subjectivity values and the y-axis shows the frequency of the distribution. We have constructed the histogram by considering intervals of 50.

## 7. Discussion

This research has focused on extracting tweets related to farmers' protest in India to analyze the sentiments. Indian farmers' protest that began in November 2020 has gained widespread international attention and received support from people from all walks of life, directly and indirectly. Bollywood celebrities have involved themselves in the discussion on Twitter.

The data on farmers' protest continues to get generated in abundance on different social media platforms. Therefore, it is very difficult to process such huge data through conventional approaches and we need high computational facilities and approaches to process it faster.

As India is a major agricultural country with wast geographical land and varying climate, the government cannot reach farmers on a one-on-one level. Swani, R. et al. in (Aswani et al., 2019) have proposed to manage the misinformation floating on social media sites. This way the extracted sentiments from users would help the government to make the collective decisions to launch new beneficiary policies. Farmers' communities depend on the policies framed by the government. Also, it is the responsibility of the governing body to provide food supply to each

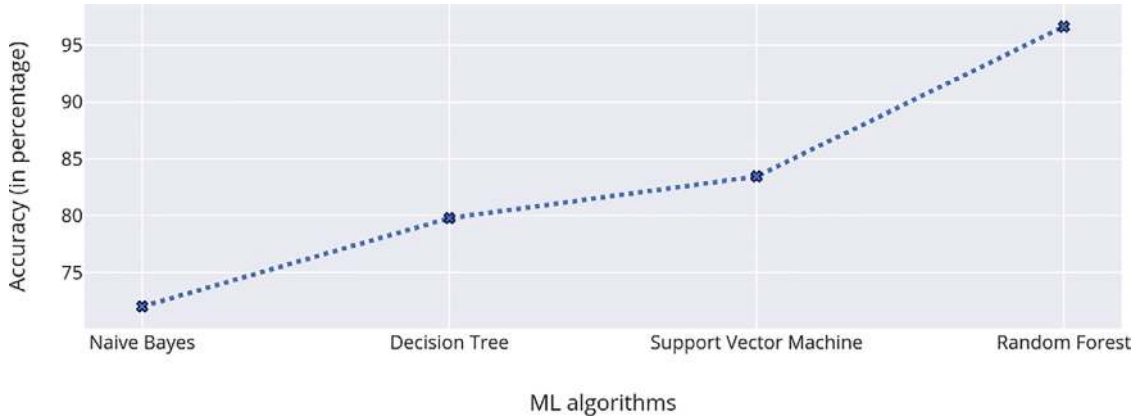


**Fig. 9.** Accuracy Curve of ML algorithms.

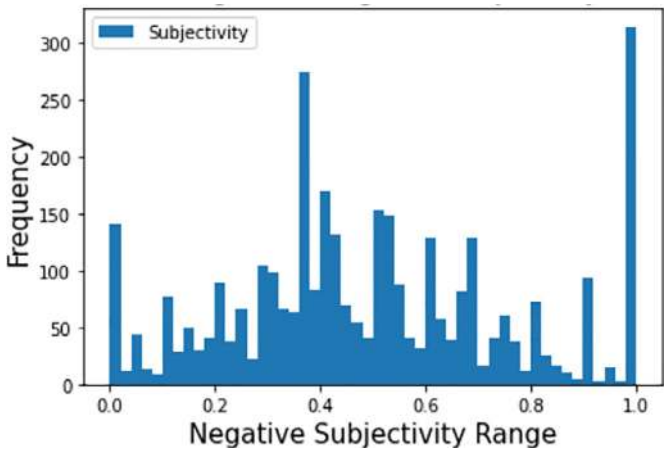


**Fig. 10.** Frequency distribution of Negative Subjectivity Range.

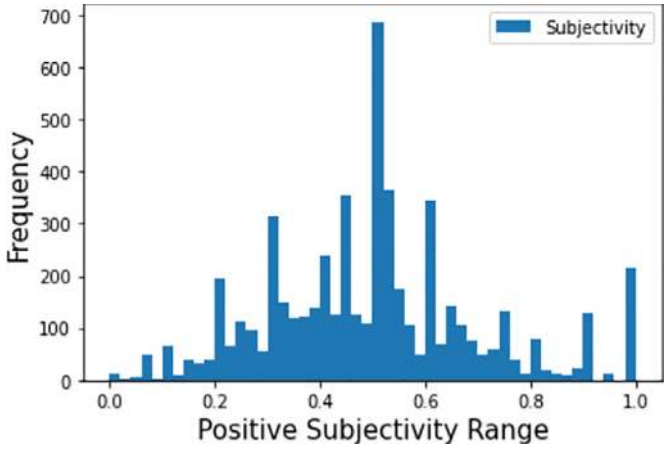


**Fig. 11.** Frequency distribution of Positive Subjectivity Range.

citizen and keep track of all information on social media to hear the difficulties faced by the farmers.

### *7.1. Contributions to literature*

In the current times, most political leaders are available on social media platforms. Authors in Grover et al. (2021) used Twitter data and analyzed the cohort-specific prioritization of the leadership which could be beneficial to society. With the low-cost availability of smartphones, most users use social media platforms to share their insights. Authors in Sarin et al. (2021) used twitter-based mining for generated content which will help to manage the information and to extract some usual information from it. Many studies have been done to classify the sentiments of social media users because the information being shared online can be viral very rapidly (Joseph et al., 2021).

Our work focuses on understanding different sentiments using multiple machine learning techniques to analyze the tweets and classify the polarity and subjectivity. We applied machine learning techniques with two different factorization techniques Bag of Words and TF-IDF and found that Bag of Words gives better accuracy than TF-IDF.

### *7.2. Implication to practice*

With the ever-increasing social media data in today's age, particularly on Twitter, there is a dire need to analyze tweets systematically and effectively. This research will be valuable to local and central governments, who are the driving forces behind the amendment of laws and regulations, as it will give a macro-level understanding to them regarding the situation at the ground level. This research will allow the government to be better prepared to handle such situations.

## 8. Conclusion and future scope

Digital platforms have given us opportunities to share our thoughts, ideas, and opinions (Sutherland & Jarrahi, 2018). Not only for this but also for propagating ideas and forming personal opinions, social networks have grown in popularity. Analyzing the details of social media sites will provide one with a perspective on culture and the environment. Due to this, The farmers' protest in India saw a humongous rise in the number of tweets where users shared their thoughts. The farmers' protest in India has created every category of people expressing their agitation towards the issue. In this paper, we have explored ways to understand the sentimentality of people by building a sentiment analysis model and identifying the direction the protest is leading towards. We discovered that the bulk of tweets are neutral, with positive sentiments coming in a close second and negative sentiments coming in last. Furthermore, four common machine learning models were used for classification and prediction. We saw that random forest yielded the best result. The keywords used to find material relevant to the farmers' protest are one of the work's limitations. If the keywords were not used in the messages, some relevant tweets were likely skipped.

One of the drawbacks of our study is that we could have extracted a comparatively large number of tweets given that millions of people expressed their opinions about the protests. More number of tweets might have been resourceful in uncovering a rather large number of

sentiments, but we lacked the computational resources to process such a huge amount of tweets.

Future research may look at various algorithms, including using unsupervised learning as the primary method, and see if the outcomes vary (Pandarachalil, Sendhilkumar, & Mahalakshmi, 2015). The study can also be expanded to look at how Covid-19 rampaged throughout India as a result of mass protests and rallies, as Covid-19 was at its peak during the months of protest.

Iwendi et al. (2020b) has suggested the use of a semantic privacy preserving framework for unstructured medical datasets. It is essential that such machine learning models which allow us to use sentiment analysis be used responsibly and the public emotions are not used to create havoc in society. It is therefore necessary that farmers' and common citizens' privacy be preserved when dealing with data.

A wide number of Twitter users expressed concern that the farmers' protest may be a result of an international propaganda to destabilise the government. M. Mittal et al. discussed an approach in Mittal, Saraswat, Iwendi, and Anajemba (2019) to detect intrusion in wireless sensor network systems. This approach can further be worked upon and implemented on a wider scale by government to ensure the democratic strength of the country.

This study might be valuable in the field of public policy, where governments may use machine learning techniques to improve decision-making and conduct mass behavioral analysis (Iwendi et al., 2020a). The study can also be beneficial to analyze multiple protests happening internationally with a more inclusive approach to various local languages.

Figure 6 shows the range of polarity of each tweet in the dataset. There are approximately 17,500 tweets in the dataset and the graph shows the range of sentiments shown by these

According to Fig. 2, the average neutral sentiment is expressed in as much as 8000+ tweets, thus ranking highest among all sentiments. Approximately 46% of the users maintained unbiased opinions of the farmers' protests. They did not support any particular entity, farmers, or the government. The positive sentiment is expressed in 5000+ tweets and the negative sentiment is expressed in approximately 3000+ tweets. 29% users expressed positive sentiments about the protest either in the form of supporting farmers or in the form of understanding the government's actions. 17% of the users used negative language while discussing farmers' protest on the microblogging website.

## References


Abdul-Mageed, M., & Diab, M. (2014). SANA: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 1162–1169).

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management, 39*(1), 45–65. 10.1016/S0306-4573(02)00021-3.

Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: Managing misinformation in social media-insights for policymakers from twitter analytics. *Journal of Data and Information Quality (JDIQ), 12*(1), 1–18.

Bakshi, R. K., Kaur, N., Kaur, R., & Kaur, G. (2016). Opinion mining and sentiment analysis. 978-9-3805-4421-2/16.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*(2), 19–227. 10.1007/s11749-016-0481-7.

Bugden, D. (2020). Does climate protest work? Partisanship, protest, and sentiment pools. *Socius: Sociological Research for a Dynamic World, 6*. 10.1177/2378023120925949. P. 237802312092594

Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. *Proceedings of the National Conference on Artificial Intelligence, 2*, 1515–1521.

Chintalapudi, N., Battineni, G., Canio, M. D., Sagaro, G. G., & Amenta, F. (2021). Text mining with sentiment analysis on seafarers' medical documents. *International Journal of Information Management Data Insights, 1*(1), 100005. 10.1016/j.jjimei.2020.100005.

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing, 101*, 107057. 10.1016/j.asoc.2020.107057.

Go, Huang, & Bhayani (2009). Twitter sentiment analysis (final project results). *Journal of Information*.

Grover, P., Kar, A. K., Gupta, S., & Modgil, S. (2021). Influence of political leaders on sustainable development goals-insights from twitter. *Journal of Enterprise Information Management*.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In *Proceedings of the... annual hawaii international conference on system sciences. Annual hawaii international conference on system sciences* (pp. 1833–1842). 10.1109/HICSS.2014.231.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 17*(3), 299–310. 10.1109/TKDE.2005.50.

Iwendi, C., Khan, S., Anajemba, J. H., Mittal, M., Alenezi, M., & Alazab, M. (2020a). The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems. *Sensors, 20*, 2559. 10.3390/s20092559.

Iwendi, C., Moqqurab, S. A., Anjum, A., Khan, S., Mohan, S., & Srivastava, G. (2020b). N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets. *Computer Communications, 161*, 160–171. 10.1016/j.comcom.2020.07.032.

Jain, S., Seeja, K. R., & Jindal, R. (2021). A fuzzy ontology framework in information retrieval using semantic query expansion. *International Journal of Information Management Data Insights, 1*(1), 100009. 10.1016/j.jjimei.2021.100009.

Jain, T. I., & Nemade, D. (2010). Recognizing contextual polarity in phrase-level sentiment analysis. *International Journal of Computers and Applications, 7*(5), 12–21. 10.5120/1160-1453.

Jivani, A.. A comparative study of stemming algorithms. International Journal of Computer Technology and Applications, 2(6), 1930–1938.

Joseph, N., Kar, A. K., & Ilavarasan, P. V. (2021). How do network attributes impact information virality in social networks? *Information Discovery and Delivery*.

Kotsiantis, S. B., & Kanellopoulos, D. (2006). Data preprocessing for supervised learning. *International Journal of Computer Science, 1*(2), 1–7. 10.1080/02331931003692557.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the OMG!ICWSM,.

Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021). Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights, 1*(1), 100008. 10.1016/j.jjimei.2021.100008.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Association of Computational Linguistics*, 142–150.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The stanford coreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60).

Mihalcea, R., Banea, C., & Wiebe, J. (2010). Multilingual sentiment and subjectivity analysis. In *Multilingual natural language processing applications: From theory to practice* (pp. 1–19).

Mishra, R. K., Urolagin, S., & Jothi, A. A. J. (2019). A sentiment analysis-based hotel recommendation using TF-IDF approach. In *2019 International conference on computational intelligence and knowledge economy (ICCIKE)* (pp. 811–815). 10.1109/ICCIKE47802.2019.9004385.

Mittal, M., Saraswat, L. K., Iwendi, C., & Anajemba, J. H. (2019). A neuro-fuzzy approach for intrusion detection in energy efficient sensor routing. In *2019 4th International conference on internet of things: Smart innovation and usages (IoT-SIU)* (pp. 1–5). 10.1109/IoT-SIU.2019.8777501.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics, 18*(6), 275–285. 10.1002/cem.873.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. 1-58113-583-1/03/0010.

Nithya, V. I. (2016). Preprocessing techniques for text mining. Vol. 5, no. October 2014, pp. 7–16.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the 7th international conference on language resources and evaluation, 2010, 1320–1326. 10.17148/ijarcce.2016.51274

Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. S. (2015). Twitter sentiment analysis for large-scale data: An unsupervised approach. *Cognitive Computation, 7*, 254–262. 10.1007/s12559-014-9310-z.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. 10.3115/1218955.1218990.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. 10.3115/1118693.1118704.

Pietra, V. J. D., Berger, A. L., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Association for Computational Linguistics*.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics, 3*(2), 143–157. 10.1016/j.joi.2009.01.003.

Rajman, M., & Besancon, R. (1998). Text mining: Natural language techniques and text mining applications.

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights, 1*(2), 100012. 10.1016/j.jjimei.2021.100012.

Sarin, P., Kar, A. K., & Ilavarasan, V. P. (2021). Exploring engagement among mobile app developers-insights from mining big data in user generated content. *Journal of Advances in Management Research*.

Sarlan, A., Nadam, C., & Basri, S. (2014). Twitter sentiment analysis. In *Proceedings of the 6th international conference on information technology and multimedia* (pp. 212–216). 10.1109/ICIMU.2014.7066632.

Soomro, Z. T., Ilyas, S. H. W., & Yaqub, U. (2020). Sentiment, count and cases: Analysis of twitter discussions during COVID-19 pandemic. In *2020 7th International conference on behavioural and social computing (BESC)* (pp. 1–4). 10.1109/BESC51023.2020.9348291.

Srivastava, A., Singh, V., & Drall, G. S. (2019). Sentiment analysis of twitter data: A hybrid approach. *International Journal of Healthcare Information Systems and Informatics, 14*(2), 1–16. 10.4018/IJHISI.2019040101.

Sutherland, W., & Jarrahi, M. H. (2018). The sharing economy and digital platforms: A Review and research agenda. *International Journal of Information Management, 43*, 328–341. 10.1016/j.ijinfomgt.2018.07.004.

Swain, P. H., & Hauska, H. (1977). Decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics, 15*(3), 142–147. 10.1109/tge.1977.6498972.

Szabolcsi, A. (2004). Positive polarity - negative polarity. *Natural Language & Linguistic Theory, 22*(2), 409–452. 10.1023/B:NALA.0000015791.00288.43.

Webster, J., & Kit, C. (1992). Tokenization as the initial phase in NLP.

Yang, F. (2018). An implementation of naive bayes classifier. In *2018 International conference on computational science and computational intelligence (CSCI)* (pp. 301–306). 10.1109/CSCI46756.2018.00065.

Zervoudakis, S., Marakakis, E., Kondylakis, H., & Goumas, S. (2021). Opinionmine: A Bayesian-based framework for opinion mining using twitter data. *Machine Learning with Applications, 3*, 100018. 10.1016/j.mlwa.2020.100018.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories Technical Report 89*.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics, 1*(1–4), 43–52. 10.1007/s13042-010-0001-0.