SENTIMENT ANALYSIS AND CLSSISIFACTION OF INDIAN FARMERS PROTEST USING TWITTER DATA

Dr. Rajendra Dev[1], Matta Hema[2], Matcha Chandana[3], Penumathsa Jyotsna[4], Nagala Tejaswini [5]

Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

ABSTRACT

Protests are essential to democracies because they allow citizens to voice their needs and/or unhappiness to the government. As people become more aware of their rights, there have been increasing protests across the globe for a variety of causes. Together with technical advancement, social media usage for information and idea sharing has drastically expanded. For this study, we gathered data from the micro-blogging website Twitter in order to better understand how the general public felt about the farmers' protest. We used models to categorise and evaluate the feelings based on a set of roughly 20,000 tweets regarding the protest. In our investigation, we used the TF-IDF and the Bag of Words, and in our inquiry, Bag of Words and TF-IDF were both used, and the outcomes revealed that Bag of Words performed better than TF-IDF. Also, we used Decision Trees,Support Vector Machines, Naïve Bayes. Text blob is used to analyse the data.

INTRODUCTION

In India's northern regions, farmers are still protesting against the three farm acts that Parliament enacted in September 2020. 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 are the three acts in question. These Acts will transform Indian agriculture according to the government and "attract private investment". Contract farming is made possible by the Farmers' (Empowerment and Protection) Agreement on Price Assurance and Farm Services Act, 2020. In this system, farmers produce crops in exchange for an agreed-upon payment under agreements with corporate investors. Farmers who are protesting fear that wealthy investors would obligate them to unfair contracts drafted by significant corporate legal firms, subjecting them to liabilities.

The three acts will be recalled as per their demand, and more than 40,000 demonstrators have vowed to ensure it happens. The government's suggestion that the regulations be suspended for 18 months was rejected by the farmers, and the government has maintained that the protests are the product of false information.

Protests are an essential component of a democratic society and they have the power to significantly influence how that society develops in the future. The entire society steps forward to express its ideas in favour of or against a community when it protests. This is important for the growth of society. Many of these demonstrations have helped to remove long-held views that were no longer valid in the modern world. The protests also give the common people a chance to speak with their elected representative. Nonetheless, some forms of protest can sometimes result in violence and upset the social order. Understanding the feelings behind online discussions is crucial to comprehending a protest because it enables us to include a wider audience and include both direct and indirect participants.

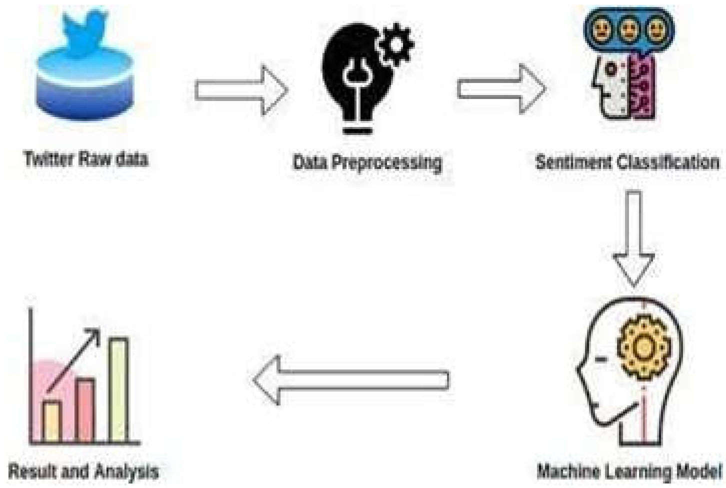
As a result of the widespread protest, there have been a lot of opinions and feelings expressed on social media on a global scale

As a result of the huge protest. The general public's attitudes and opinions covered a wide spectrum. Thousands of individuals shared their idea son social media in support of the farmers, who were backed by people all over the world. The following hash tags were trending on

Twitter: #farmers protest, #i am with farmers, #Speak Up For Farmers, #IStand With Farmers, and #kisan ekta zindabaad. Many organisations, never the less, spoke up and dubbed it anti-national ropaganda as well. Several groups also held the opinion that the protests were taking place due of a lack of knowledge and that the farm legislation were in the farmers' favour.

METHODOLOGY

Here,we go through the process we used to research public opinion on the on-going farmers' protest in India. The methodical procedures we used to evaluate and forecast the sentiment of a specific Twitter user are showninFig.1. The procedure begins with gathering data from Twitter, which is followed by a few important steps including cleaning and preparing the data to convert it into a format that computer scan interpret.Then extstepistodetermineandcategoriseauser'ssentimentusingtwocriteria.Moreover, visualisation techniques can beutilised to examine how people feel about certain issues.



DATASET

In total, thousands of tweets were gathered over the course of four months.We gathered the raw data by directly accessing the Twitter API using the open-source Python module .

Link: Farmers-protest-tweet-analysis/tweetsfarmbills.csvat main ·AlbusDracoSam/Farmers-protest-tweet-analysis· GitHub

SAMPLE:

	Unnamed: 0	Date	Tweet	Tweet Source	Retweets	Likes	Location	Clean Tweet	SS
2	2	2021-01-01 20:10:18	Half the problem got sorted when @RahulGandhi ...	Twitter Web App	4	8	New Delhi, India	half the problem got sorted when left the rema...	NEGATIVE
9	9	2021-01-01 18:10:24	Press conference by Kisan Andolan representati...	Twitter for iPhone	0	0	New Delhi, India	press conference kisan andolan representatives...	POSITIVE
16	16	2021-01-01 16:41:18	@narendramodi @PMOIndia @nstomar \nThis is all...	Twitter Web App	0	1	New Delhi, India	this all you people and never forget that hist...	NEGATIVE
55	55	2021-01-01 04:46:15	Watch here farmers protests full documentary\n...	Twitter for Android	0	0	New Delhi, India	watch here farmers protests full documentary c...	NEGATIVE
57	57	2021-01-01 03:51:31	@mastersheefu_\n\nGood Morning\n 🙏 🙏 🙏 \n\nTo supp...	Twitter for Android	1	2	New Delhi, India	good morning support farmers posting this picture	POSITIVE
...	...	...	...	...	...	...	...	...	...
3359	3359	2020-12-18 09:26:49	"New Agri laws haven't come overnight..."\n\nR...	Twitter Web App	1	1	New Delhi, India	new agri laws havent come overnight read what ...	NEGATIVE
3364	3364	2020-12-18 09:18:22	The Farmers Acts – A Bird's Eye View. Written ...	Twitter for Advertisers (legacy)	0	3	New Delhi, India	the farmers acts birds eye view written adv ma...	POSITIVE

DATA PRE-PROCESSING

Pre-processing the data is absolutely necessary and calls for a method called data cleaning, which entails converting raw data into a format that can be understood by machines. We have a sizable text dataset made up of tweets, so we need to clean it to get rid of certain discrepancies in order to prevent inconsistent data. We use a rather straight forward method to sanitise the data.Before we could begin cleaning up the data, we used Excel's built-in feature to get rid of any duplicate records that were there.
Using regular expressions, we started by deleting@-mentions,Re-tweets denoted by the letter "RT," links, and hash-tag symbols because they don't offer any value. It's important to note that we took care to leave the phrases that follow the hash-tag in place because they can provide important context for the tweet's tone. For instance,theline"Istandwithfarmers"withthehash-tag#istandwithfarmerprovidesus with information on the user's mental state despite the fact that the symbol '#' has no impact on our analysis either positively or negatively. Moreover, tweets with punctuation, emojis, special characters, and digits were deleted.The term "tokenization" refers to the process of breaking up large amounts of text into tokens. In order to model text data,tokenization is a crucial step. Analysing the word order facilitates comprehension of the text's meaning. To lessen the inflection towards their root forms, we used a porter stemmer. The suffix was stripped away to create stems in order to do this. Finally, we added a new Pandas column called "Cleaned Tweets" to our current data frame containing the tweets dataset to store the fully pre-processed tweets.

Lexicon based sentiment calculation

In our project, we have chosen to use a lexicon-based strategy in order to avoid the process of creating labelled data. Calculating the semantic orientation of words found in the text is necessary for this strategy. A lexicon-based approach has the main benefit of being considerably easier to grasp and modify by a person. The semantic orientation can be identified and classified as neutral, positive, or negative using this method. Semantic orientation assesses the polarity and strength of the text, whereas sentiment analysis is defined as a method to extract subjectivity and polarity from text. The sentiment orientation value is then calculated using various adverb and adjective combinations. Furthermore, the entire value is available from a single source. Text-Blob, a well-known Python library, is used, which provides intricate operations and analysis on the tweet data. The tweets are presented in a numerical style, and Text-Blob gives each tweet a unique score. Last but not least, a pooling procedure that averages all the sentiments in a tweet determines its sentiment.

Sentiment classification

The polarity and subjectivity of the tweet are the two values that Text-Blob returns, which we also need to be aware of. Polarity is defined as [-1, 1] where [1] denotes a good sentiment and [-1] a negative sentiment. Negative words flip the polarity, bringing it below zero. The range of subjectivity is [0, 1]. The tweet's subjective- ness reveals how much factual and personal details are included. When subjectivity is strong, there are more individual opinions present. Intensity is a further setting for Text-Blob. Text-Blob uses the intensity to determine subjectivity. Whether a word has any kind of influence on the next word depends on its intensity. Adverbs are employed as modifiers in English, such as "very good," which was described as using a lexicon-based approach in the preceding section. There are instances where the values are precisely 0. A sentiment score will be assigned based on the polarity value, and the computation is done in such a way that if the score is less than 0, the sentiment will be returned as negative. The sentiment is returned as positive if the polarity is higher than 0. In all other circumstances, the sentiment is considered neutral and the score is set to 0. After classifying the sentiments, we took a look at the count of the number of sentiments and found that a large number of people have neutral feelings about the protest indicating that neither they support the farmers' protest nor they support the government.

Model building

This section discusses the categorization and prediction of tweet sentiment using Naive Bayes, Decision

Tree, Random Forest, and Support Vector Machine, four well-known supervised machine learning methods. Since text data cannot be processed by computers in its raw state, the data must be cleaned up before machine learning models are trained on it. It is necessary to manually break down the language into a numerical structure that the machine can comprehend (Rawat Kumar, & Sabitha, 2021). Thus, we look at the outcomes of two NLP techniques: Bag of Words and Term Frequency and the Inverse Document Frequency Approach (Zhang, Jin, & Zhou, 2010). The NLP algorithms BoW and TF-IDF enable us to transform tweet phrases into numerical vectors (Manning et al., 2014).

Bag of words and term frequency-inverse document frequency

Bag of words

As in our example of machine learning algorithms for tweet sentiment categorization, the Bag of Words model isa method of extracting features from a text that can be utilised in modelling. Simply said, it is a collection of words that are used to describe a sentence in a document that has a word count.A list of well-known phrases is the first component, and a metric for assessing their presence is the second. The order in which they occur is another characteristic of BoW. The first step is to create a vocabulary from each unique word in our data frame of tweets. List each of these distinctive words separately and keep track of their presence in each and every tweetas the following step.

TF-IDF

Due to its use in determining the significance of a word in a tweet, the TF- IDF technique performs better than the BoW approach. When evaluating word frequency, one typical problem is that frequently occurring terms start to take over the text yet may not have the necessary "informational content" for the model to distinguish between them appropriately. A metric for assessing a word's importance is the IDF. Since determining the TF alone is insufficient to understand the meaning of words, we require the IDF value: The calculation of term frequency for the term t in document d is shown in Eq. (1). The word frequency score is based on how often a term appears in the document.

Eq.(1) $TF(t,d)= N(t,d) /T$

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.

Eq.(2) $IDF(t)=\log N/(N(t))$

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

Eq.(3) $TF\text{-}IDF=TF*IDF$

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

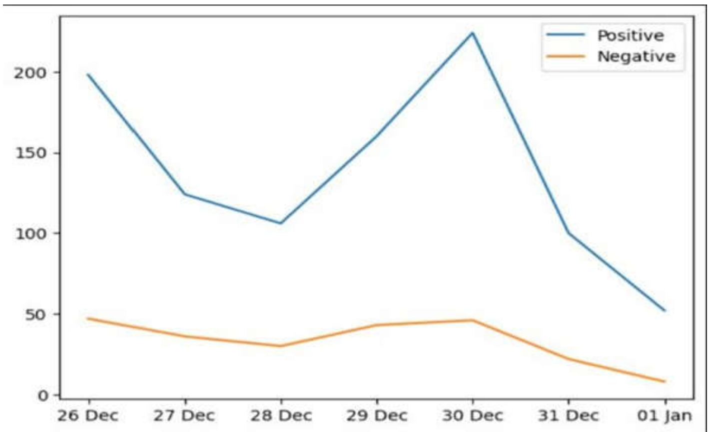
Text Blob

A Python library for Natural Language Processing is called Text-Blob (NLP). Natural Language Tool-Kit (NLTK) was a tool that Text-Blob actively employed to complete its tasks. The NLTK library enables users to do categorization, classification, and a variety of other tasks while providing quick access to a large number of lexical resources. Text-Blob is a straightforward library that provides intricate textual analysis and processing.

A sentiment is identified by its semantic orientation and the force of each word in the sentence for lexicon-based techniques. This calls for a pre-defined dictionary that divides words into negative and positive categories. A text message will typically be represented by a bag of words. Following the individual scoring of each word, the ultimate sentiment is determined by performing a pooling procedure, such as averaging all the sentiments.

Experimental results and analysis

The subjectivity is dispersed throughout the graph, whereas the polarity is primarily concentrated in the middle. This suggests that there is a large range of subjectivity in our collection of tweets, and the most of them fall in the [-0.75, 0.75] polarity range, suggesting that the extremes of negative or positive emotion are noticeably low. While users have expressed their full opinions as well as information about the farmers' protest online, the majority of tweets display a range of emotions, both good and negative. Only a very small percentage of users have shown their support for the farmers' protest in violent terms. Tweets with low subjectivity are clustered in the middle of the polarity range [-1, +1] in the graph, whereas tweets with high subjectivity are dispersed throughout the polarity range [-1, +1]. This suggests that while the facts surrounding the farmers' protest were neutral, the users' expressed opinions ranged widely from negative to favourable. This makes sense given that a fact (low subjectivity) is more likely to be neutral (with polarity 0) and an opinion (high subjectivity) is more likely to have a wide range of emotions, from negative to positive.



CONCLUSION

We now have the ability to express our thoughts, ideas, and opinions through digital media. Social networks have gained popularity not just for this but also for spreading ideas and creating personal beliefs. One can gain perspective on society and the environment by looking at the specifics on social media platforms. As a result, there was a tremendous increase in the quantity of tweets from individuals who expressed their opinions about the Indian farmers' protest. Every type of person has shown agitation over the issue as a result of the Indian farmers' protest. By developing a sentiment analysis model and determining the direction the protest is heading, we have investigated methods for understanding the sentimentality of people in this study.

REFERENCES

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

2. 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.

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

4. G. Biau, E. Scornet. A random forest guided tour

5. E. Cambria, D. Olsher, D. Rajagopal. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis

6. N. Chintalapudi, G. Battineni, M.D. Canio, G.G. Sagaro, F. Amenta. Text mining with sentiment analysis on seafarers’ medical documents

7. K. Garcia, L. Berton. Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA

8. Go, Huang, Bhayani. Twitter sentiment analysis (final project results)

9. P. Grover, A.K. Kar, S. Gupta, S. Modgil. Influence of political leaders on sustainable development goals-insights from twitter

10. F. Heimerl, S. Lohmann, S. Lange, T. Ertl Word cloud explorer: Text analytics based on word clouds

11. J. Huang, C.X. Ling. Using AUC and accuracy in evaluating learning algorithms

12. C. Iwendi, S. Khan, J.H. Anajemba, M. Mittal, M. Alenezi, M. Alazab. The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems

13. T.I. Jain, D. Nemade. Recognizing contextual polarity in phrase-level sentiment analysis