Social Media Analytics (SMA) enabled by Text Mining : *Farmer Protests in India*

Aryaan Shome
MSc Advanced Computer Science, Department of Computer Science
University of Manchester
Manchester, UK
Student ID: 10731392

Abhinav Thomas
MSc Advanced Computer Science, Department of Computer Science
University of Manchester
Manchester, UK
Student ID: 10781010

Anusha Haleema
MSc Advanced Computer Science, Department of Computer Science
University of Manchester
Manchester, UK
Student ID: 10774080

Regson Do Rego
MSc Advanced Computer Science, Department of Computer Science
University of Manchester
Manchester, UK
Student ID: 10780106

Shefali Singhal
MSc Advanced Computer Science, Department of Computer Science
University of Manchester
Manchester, UK
Student ID: 10756907

Shruti Govind
MSc Advanced Computer Science, Department of Computer Science
University of Manchester
Manchester, UK
Student ID: 10763314

***Abstract*— Computers always found it difficult to understand the natural human language until the advent of automated analysis of data mined from text-based datasets. Text analysis is a process which is employed after text mining to derive high quality information from the text. Today, text analysis is often used in conjunction with glossaries containing words pertaining to human sentiments to analyze sentiments of people on social media platforms such as Twitter. This provides an efficacious method to ascertain the public opinion on a topic of interest. Most of the tools used for sentiment analysis require the glossary to automatically classify sentiments present in a piece of text. In this project a social media analysis was carried out on Twitter feeds with the aim to analyze the emotions and opinions of people towards the farmer protests in India. This research paper depicts the methodology followed for sentiment analysis and the results obtained on successfully running the code.**

I. Introduction

Social Media is defined as "web-based and mobile-based Internet applications that allow the creation, access and exchange of user-generated content that is ubiquitously accessible" [1]. At a time when social media is burgeoning, it is imperative that we avail of the current technologies to extract as much useful information as we can from it. Social media is like a repository of popular opinions and sentiments. It is meaningful to utilize the massive datasets available as feeds on Twitter to make informed decisions when it comes to issues affecting people directly. Now, humans are not required to manually go through each tweet and record the sentiments. Text mining has made it easier to parse enormous amounts of textual data and retrieve relevant information. Data scientists and researchers are giving much attention to natural language processing using text analysis. Sentiment analysis has taken text analysis to a whole new level of advancement in technology. In this research paper, we proffer a futuristic application of sentiment analysis on the opinions of people from around the world regarding the farmer protests in India using data obtained from Twitter API. We processed the data before analyzing it, made use of topic modelling and implemented Named Entity Recognition (NER)/Named Entity Linking (NEL) to arrive at the results. At the end of the paper, we show our analysis of the result-set.

II. Related Work

A. Understanding Text Mining

"Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections."[2]

There are five main techniques of text mining:

- **Information Extraction (IE)** – This technique processes small and large pieces of semi-structured and unstructured texts to identify attributes, entities, and the relationships between them.
- **Information Retrieval (IR)** – Relevant patterns are extracted from the text by giving primary attention to words and phrases provided to the analyzer. It is used to monitor user behavior.
- **Categorization** – It employs supervised learning where texts in human language are assigned to a set of topics which has been defined already. The

correct topics and indexes for the documents are determined using Natural Language Processing.

- **Clustering** – Hidden patterns and structures are identified in texts and then organized into pertinent clusters. A major challenge is to form meaningful clusters out of unlabeled data.
- **Summarization** – It is one of the most useful techniques in text mining. The miner browses through multiple documents to weave a summary containing a commendable amount of important information while simultaneously maintaining brevity and holding up the overall intent of the original documents.

B. Social Media Analysis Overview

In this paper we group the whole process of analyzing the data obtained from Twitter API into four steps –

- **Data Preprocessing** – It is the method which is used to convert raw unstructured or semi-structured data into a format that is understandable by the computer. This helps to fill in the missing values, do away with discrepancies and get rid of noisy data, often containing outliers.
- **Topic Modelling** – A statistical modelling technique which employs unsupervised machine learning. Without using a lexicon of keywords or regular expressions, it scans documents, detects associated patterns, and uncovers the abstract topics appropriate for forming groups of these documents.
- **Sentiment Analysis** – Also known as opinion mining, it utilizes Natural Language Processing along with computational linguistics to classify text and determine the emotion of the textual data. It aids in gauging whether the opinions are neutral, negative, or positive.
- **Named Entity Recognition (NER)** – It falls under the Information Extraction technique and seeks to intelligently determine and classify named entities. These entities are grouped into categories which are predefined. The NER model thus helps in the extraction of useful information after intelligent classification.

III. METHODOLOGY

At the outset, when about 100,000 twitter feeds were mined, there were myriad raw tweets which contained thousands of words irrelevant to the project. Thus, the Twitter API was queried to filter out only those Tweets with the hashtag “FarmersProtest.” This data was noisy and needed preprocessing. Therefore, all the retweets and duplicate tweets were eliminated. It was ensured that all the tweets being included had at least one retweet and were created after September 2020. The data cleaning process helped remove links, timestamps, punctuations, unnecessary line breaks, apostrophes, hashtag symbols, emojis, non-English alphabets, twitter handles, leading spaces, trailing spaces, multi spaces and empty columns. Then, words from different tweets were joined together into one string and a word cloud was generated. After this, sentiment analysis was carried out on tweets which talked specifically about farmers bills.

The next important step was to identify the main concerns and topics that were being discussed regarding farmers protests. This was Topic Modelling. It used the Latent Dirichlet Allocation (LDA) algorithm in Python to spot the topics. LDA is a statistical model that enables the observations to be reasoned out by unobserved groups that in turn explains why some portions of data are similar. Bigram and Trigram models were used on the topics obtained, along with lemmatization to get rid of words that were not nouns, verbs, adverbs, or adjectives.

A corpus was created from the dictionary of topics and then this corpus was converted into a Bag of Words (BoW) representation. The Term Frequency-Inverse Document Frequency (TF-IDF) of these topics was calculated and this further helped in computing the coherence values. Then, the dominant topic and its percentage contribution in each tweet was determined. Our code returned 12 topics and the distribution of document word counts by dominant topic was reckoned. Word clouds of top n keywords in each topic were generated. Bigger and bolder words meant a higher TF-IDF. Bar graphs were generated for word count and importance of topic keywords. After computing the model perplexity and coherence score, and generating t-distributed stochastic neighbor clustering charts, the opinion of the crowd towards these topics was mined.

Finally, the celebrity tweets were taken into account and the sentiments were analyzed. A major challenge faced was that the code was unable to identify the political stand of celebrities using Named Entity Recognition (NER) even after custom training the NER model for detecting political parties and making use of Wikidata.

IV. RESULTS AND ANALYSIS

After applying the appropriate algorithms during topic modelling and sentiment analysis we determined that the sentiments of the people regarding the farmer’s bill were mostly positive. As it can be seen that coherence score increases with the number of topics, the model with the highest coherence value of 12 just before giving a major drop was chosen.

A. Sentiment Analysis of Users

From Figure 1 we can see that almost 50% of the people were in favour of the Farmer’s Protest. While 40% of the people did not agree with it, around 6% were indecisive about whether they were in support of or against the bill.

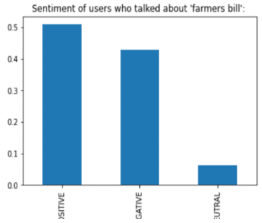


Figure 1: Sentiments of the Users regarding the Farmer’s Bill

B. *Topic Modelling*

We performed Topic Modelling to extract the hidden topics from the large volumes of data that we scraped from Twitter. Figure 2 shows a line graph of the Coherence score against the several Topics extracted from Tweets. The Coherence measures a score of a single topic by measuring the degree of semantic similarity between high scoring words in the topic.

Figure 3 shows a tabular representation of the topic numbers with their contribution percentage, keywords, and Text.

Since the coherence score seems to be increasing with the number of topics, it made sense to pick the model that gave the highest CV just before giving a major drop. In our case, this was 12.

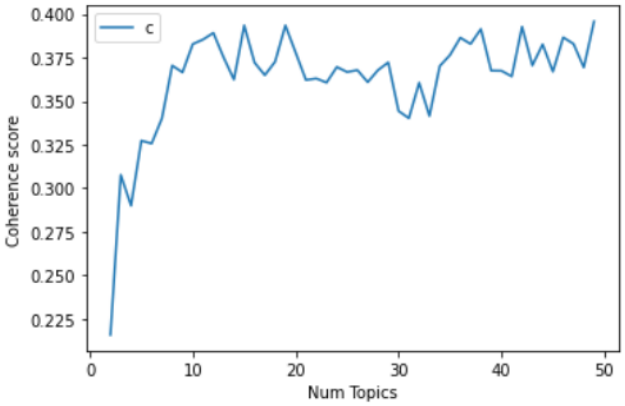
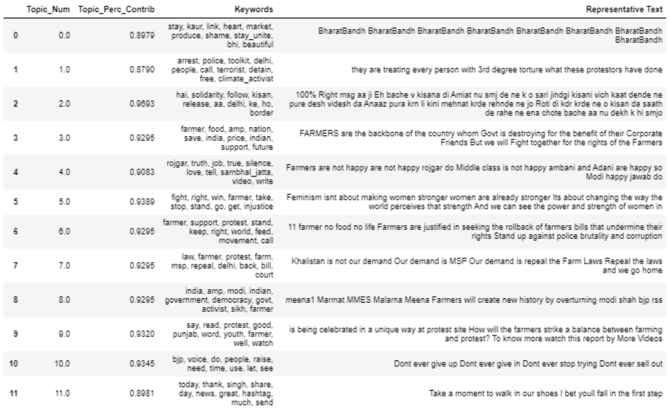


Figure 2: Line Graph depicting Coherence Scores of 50 Number of Topics

The 12 selected topics are : 1) Repeal the farmers bill - Stand with farmers 2) World Sikh day, Minimum support price - Farmers being detained 3) Write and share about human right violations 4) Toolkit for farmers bill - release the charges on the developers 5) Farmers protest reached Delhi border 6) Raise voice for Indian farmers 7) Keep supporting and sharing farmers protest - Court grant bail to arrested farmers 8) Peaceful fight against Modi government 9) Journalists and news channels supporting farmers protests 10) Farmers overcoming water shortage - which happened due to road block by gov as election closes by 11) Indian BJP government police arrest. 12) Share and support farmers - Court gives bail to farmer protesters.

For the top 12 topics selected, figure 4 shows the word cloud of top N keywords in each topic.



	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.8979	stay, kaur, link, heart, market, produce, shame, stay_unite, bhi, beautiful	BharatBandh BharatBandh BharatBandh BharatBandh BharatBandh BharatBandh BharatBandh BharatBandh
1	1.0	0.8790	arrest, police, toolkit, delhi, people, call, terrorist, detain, free, climate_activist	they are treating every person with 3rd degree torture what these protestors have done
2	2.0	0.9693	hai, solidarity, follow, kisan, release, aa, delhi, ke, ho, border	100% Right msg aa ji Eh bache v kisana di Amiat nu smj de ne k o sari jindgi kisani vich kaat dende ne pure desh videsh da Anaaz pura krn li kini mehnat krde rehnde ne jo Roti di kdr krde ne o kisan da saath de rahe ne ena chote bache aa nu dekh k hi smjo
3	3.0	0.9295	farmer, food, amp, nation, save, india, price, indian, support, future	FARMERS are the backbone of the country whom Govt is destroying for the benefit of their Corporate Friends But we will Fight together for the rights of the Farmers
4	4.0	0.9083	rojgar, truth, job, true, silence, love, tell, sambhal_jatta, video, write	Farmers are not happy are not happy rojgar do Middle class is not happy ambani and Adani are happy so Modi happy jawab do
5	5.0	0.9389	fight, right, win, farmer, take, stop, stand, go, get, injustice	Feminism isnt about making women stronger women are already stronger Its about changing the way the world perceives that strength And we can see the power and strength of women in
6	6.0	0.9295	farmer, support, protest, stand, keep, right, world, feed, movement, call	11 farmer no food no life Farmers are justified in seeking the rollback of farmers bills that undermine their rights Stand up against police brutality and corruption
7	7.0	0.9295	law, farmer, protest, farm, msp, repeal, delhi, back, bill, court	Khalistan is not our demand Our demand is MSP Our demand is repeal the Farm Laws Repeal the laws and we go home
8	8.0	0.9295	india, amp, modi, indian, government, democracy, govt, activist, sikh, farmer	meena1 Marmat MMES Malarna Meena Farmers will create new history by overturning modi shah bjp rss
9	9.0	0.9320	say, read, protest, good, punjab, word, youth, farmer, well, watch	is being celebrated in a unique way at protest site How will the farmers strike a balance between farming and protest? To know more watch this report by More Videos
10	10.0	0.9345	bjp, voice, do, people, raise, need, time, use, let, see	Dont ever give up Dont ever give in Dont ever stop trying Dont ever sell out
11	11.0	0.8981	today, thank, singh, share, day, news, great, hashtag, much, send	Take a moment to walk in our shoes I bet youll fall in the first step

Figure 3: A table showing Topic numbers with their respective contribution percentages, keywords, and Text.

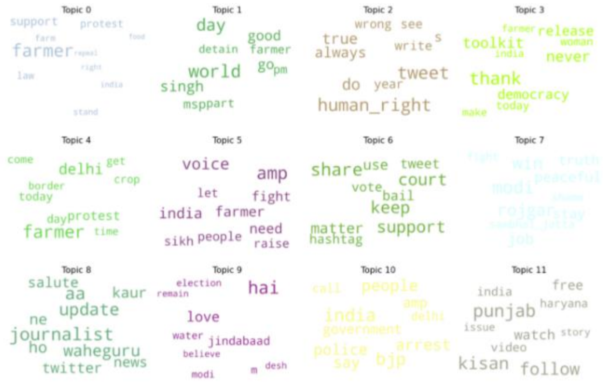


Figure 4: Word Cloud of the 12 dominant Topics

C. *Sentiment Analysis of crowd towards Dominant Topics*

Figure 5 shows the sentiments of the crowd over these 12 dominant topics. On the src -> topic-modelled.html file we can see the sentiment of each topic regarding the farmer's bill.

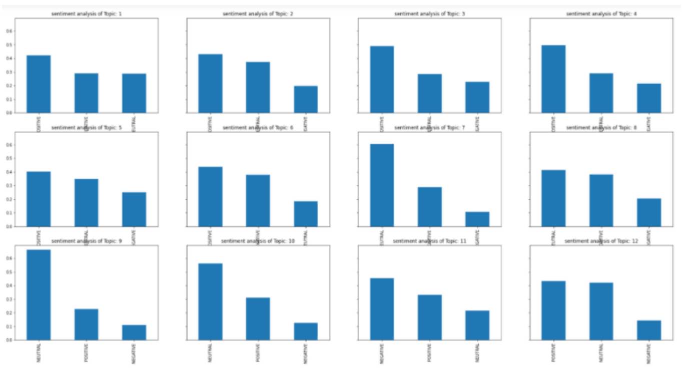


Figure 5: Sentiments of the crowd on the 12 dominant topics

D. Celebrity Tweets regarding Farmers Bill

We extracted tweets of several verified celebrity accounts on Twitter about their opinion on the Farmer's bill. Figure 6 shows their sentiments regarding the same.

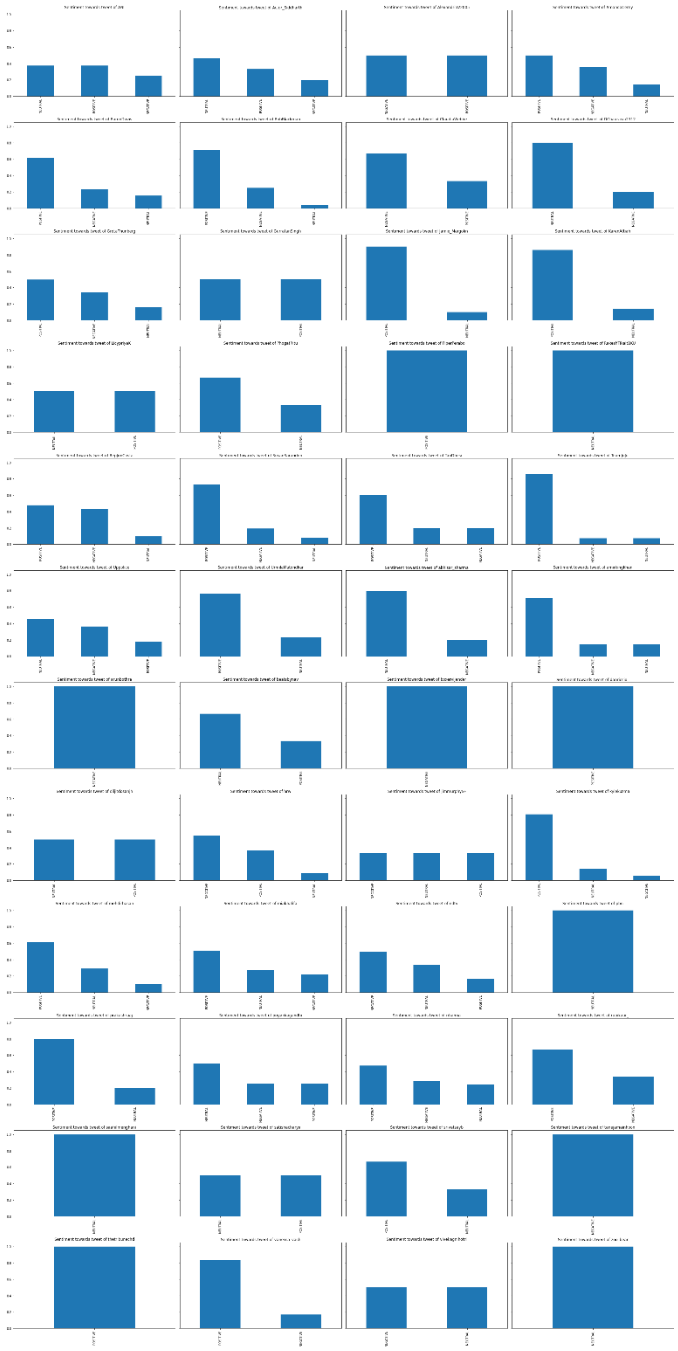


Figure 6: Sentiments of 50 celebrities on the Farmer's Bill

V. CONCLUSION

A generic sentiment analysis was carried out on the data obtained from Twitter related to the farmers protests in India. Twelve main concerns were being discussed in social media which were identified through topic modelling. The graphs depicting the results of sentiment analysis show that the sentiment of the crowd was mostly positive regarding the farmers protests. The top 50 celebrities who tweeted about it were identified as well and the sentiment of the crowd toward these celebrity tweets was summarized.

VI. REFERENCES

[1] Kaplan and Haenlein 2010.

[2] Ronen Feldman and James Sanger , The Text Mining Handbook: *Advanced Approaches in Analysing Unstructured Data*.