Universitat Pompeu Fabra Barcelona

Paula Farrás 254614
Sergi García 251425

Text Processing and Exploratory Data Analysis:

Project Part 1 IRWA

Introduction

This analysis aims to gain a better understanding of the Farmers' Protest tweets dataset by exploring key statistics and patterns in the data. The dataset consists of tweets, including metadata such as the number of likes, retweets, and URLs, along with textual content related to the Farmers' Protest in India. The exploratory data analysis (EDA) covers word frequency, tweet length distribution, hashtag usage, most retweeted tweets, and named entity recognition (NER).

The analysis utilizes various tools and libraries, such as **word clouds** for word frequency visualization, **Spacy** for named entity extraction, and **pandas** for statistical summaries.

1. Vocabulary and Tweet Content

Vocabulary Size: The dataset contains a total of **79,157 unique words**, indicating a diverse range of expressions used by Twitter users to discuss the Farmers' Protest.

Average Tweet Length:

- The average length of a tweet in terms of words is **14.98 words** per tweet, which is typical for tweets due to the platform's character limit.
- The average number of characters per tweet is **132.44 characters**, which seems short as the maximum is 280 nowadays, the question is, though, was this data extracted from Twitter when the maximum characters were 140?

Most Frequent Words: The most common words in the dataset, as expected, revolve around key themes and hashtags related to the protest:

- "**#farmersprotest**" is the most frequent term, appearing **49,819 times**. We can notice how this term is a hashtag, this further proves the relevance of a hashtag in a tweet. We will see a deeper analysis on hashtags later on this report.

Other frequent terms include:

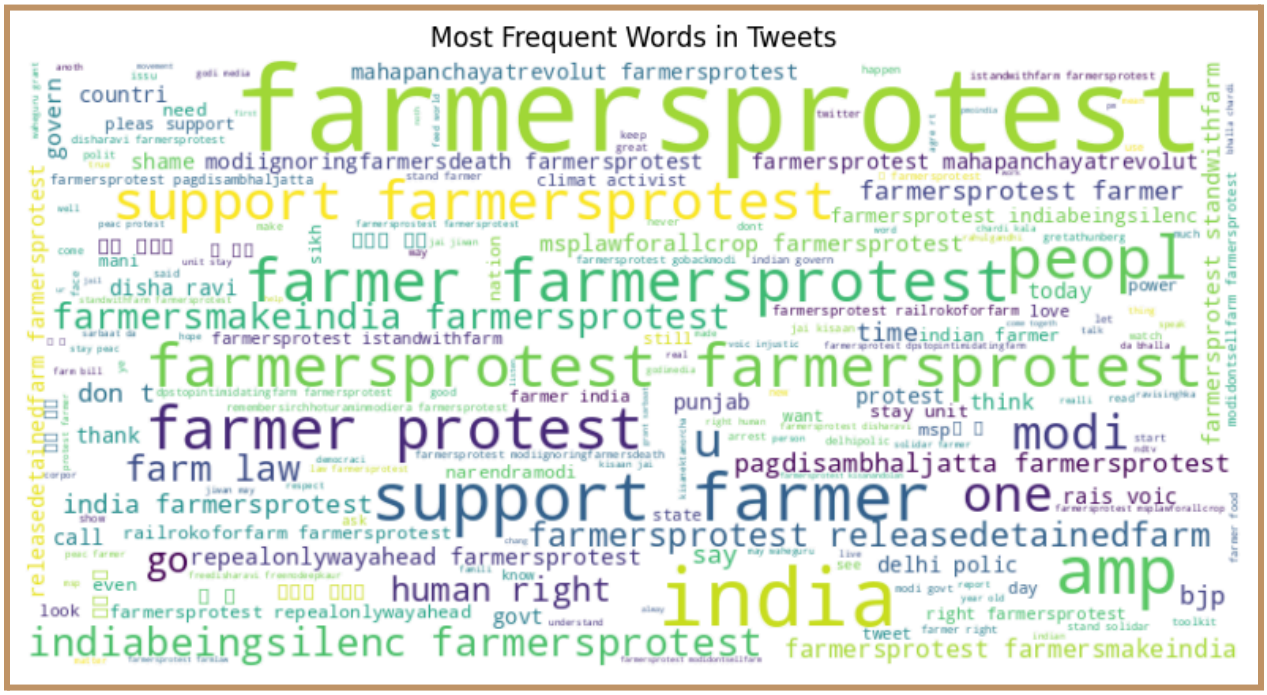
Paula Farrás 254614
Sergi García 251425

- "farmer" (15,333 occurrences)
- "India" (6,192 occurrences)
- "support" (5,953 occurrences)
- "protest" (4,663 occurrences)

These words indicate the central focus of the discussions — supporting farmers, addressing the Indian government, and protesting for farmers' rights.

2. Word Cloud Visualization

The **World Cloud** is used to generate a visual representation from the dataset that displays text data in nearly as meaningful ways. The **#farmersprotest** and the word "farmer" tells us these conversations are mainly for those who are in solidarity with protesters and want government action. This word cloud is a visual representation of the key subjects mentioned in this dataset.



3. Most Retweeted Tweets

A key part of understanding the dataset is identifying which tweets gained the most traction. Here are the top 5 most retweeted tweets:

Paula Farrás 254614
Sergi García 251425

TOP 5 TWEETS RETWEETED	DESCRIPTION	NUMBER RETWEETS	LIKES
TOP 1	Discusses farmers being forced to protest for their payment.	7.723	25.824
TOP 2	Criticizing Germany's non-response to farmers' protests vs. India's reaction.	6.164	27.888
TOP 3	On the arrest of Disha Ravi, a climate activist.	4.673	8.974
TOP 4	Commenting on courtroom proceedings related to the toolkit controversy.	3.742	10.403
TOP 5	A call for support by tagging influential personalities.	3.332	17.325

These retweeted posts reflect some of the core issues of the movement, such as government repression, toolkit activism, and global awareness (through influencers like Rihanna and Amanda Cerny). We can see a direct relationship between the most retweeted tweets and the most used hashtags, as in all 5 tweets, one the top 3 hashtags is used at least once.

Paula Farrás 254614
Sergi García 251425

4. Hashtag Analysis

We wanted to see the importance of the hashtags in relation to the retweets. As mentioned above, we have found a relationship. The most frequent hashtags in the dataset are:

HASHTAG	USED IN	DESCRIPTION
#farmersprotest	50,418 tweets	Signifying it as the central tag for this movement
#releasedetainedfarmers	2,439 times	Showing a significant conversation around detained farmer
#indiabeingsilent	2,146 tweets	Likely highlighting concerns over media silence or perceived indifference.

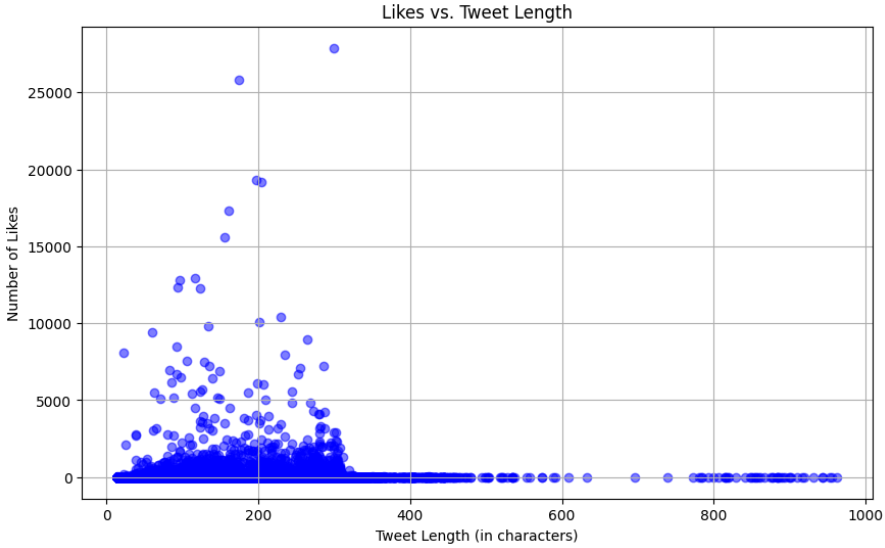
The hashtags show how the movement is not only about the protest itself but also involves global attention, human rights concerns, and calls for immediate action.

5. Visual Analysis of Data

Paula Farrás 254614
Sergi García 251425

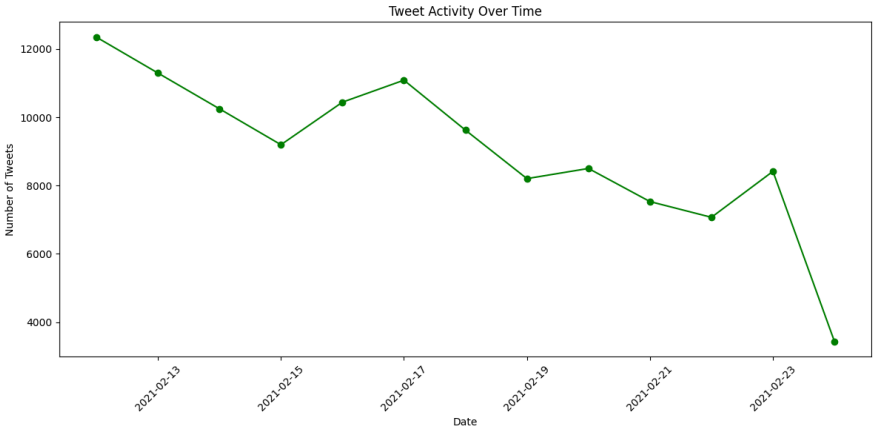
A. Scatter Plot Likes vs Tweet Length:

We can see how there is no clear correlation between Likes and Tweet Length as there are short tweets with a lot of likes and long tweets with few to none.



B. Time-based Analysis of Tweet Activity:

There is a decreasing tendency in the tweet activity from 13/02/2021 to 23/02/2021. This goes directly related to the decrease of the protests during time.



Paula Farrás 254614
Sergi García 251425

Conclusion

The Farmers' Protest dataset provides a rich understanding of the discourse on social media surrounding these protests. Using all kinds of words and repeating key hashtags brings increased visibility among both local and global audiences alike, as well as mentioning significant figures.

Analyzing hashtags of the items in retweet lists, we observe to what extent influencers, media and social justice topics dominate on world violence. Data visualization helps us understand what we are coding and analyzing but with our own eyes, not through tables and lines of code.

This opens the door for further research. We could also look into sentiment analysis to figure out how positive/neutral/negative the tweets are and a little more about sentiments or dive deep in topic modeling.