Alba Yerga - u198634 - 252197
Alejandro Vílchez - u189522 - 242557
Mar de la Fuente - u199328 - 253535

Report Part 1: Text Processing and Exploratory Data Analysis

GitHub repository: https://github.com/albayerga/G_102_6

1. Preprocessing

This report analyzes the hashtag usage within a collection of tweets, focusing on the patterns and implications of their composition. Upon examining the original data, we observe that many hashtags feature capitalized words, exemplified by hashtags such as #ModiDontSellFarmers and #FarmersProtest. For the purposes of this analysis, we will assume that a hashtag like #FarmerProtest is treated as equivalent to the phrase "farmer protest." This approach allows us to better understand the context and sentiment surrounding these discussions.

In order to implement the previous assumptions we have defined this two functions:

- **Build terms** This function processes a given text line by t removing any URLs from the text and tokenizes the line into individual words. If a word is a hashtag, it separates it into individual words, assuming each separate word starts with a Capital. The function converts all words to lowercase and removes punctuation. It filters out stop words and applies stemming to the remaining words, retaining only alphanumeric words. Finally, it returns a list of clean and processed terms

For example: build_terms("hello my #FarmersProtest is @john. I am a student, #student, ¢¢¢¢¢") would return ['hello', 'farmers', 'protest', 'john', 'student', 'student']

Given that we identified different languages among the tweets in the original data, we decided to apply stop words in the **build_terms** function according to their respective languages. We achieve this by analyzing the language column and applying the appropriate stop words for each language.

The **preprocess_document** function is designed to take a dataset of tweets as input and produce a preprocessed DataFrame as output. It begins by creating a copy of the original dataset to avoid any modifications to the original data. The function then iterates through each tweet in the DataFrame, retrieving the content and the corresponding language for each tweet. It checks if the tweet's language is present in the language_dict, assigning the appropriate language code; if the language is not found, it defaults to 'english'. The function then calls the build_terms function to preprocess the tweet content based on the identified language. After processing all the tweets, the function rearranges the DataFrame columns to a specified order: Tweet, Date, Hashtags, Likes, Retweets, and URL. Finally, it returns the preprocessed DataFrame, which now contains the cleaned tweet content organized for further analysis.

-

	Tweet	Date	Likes	Retweets	Url	Hashtags
0	The world progresses while the Indian police a...	2021-02-24 09:23:35+00:00	0	0	https://twitter.com/ArjunSinghPanam/status/136...	[#ModiDontSellFarmers, #FarmersProtest, #FreeN...
1	#FarmersProtest \n#ModiIgnoringFarmersDeaths \...	2021-02-24 09:23:32+00:00	0	0	https://twitter.com/PrdeepNain/status/13645062...	[#FarmersProtest, #ModiIgnoringFarmersDeaths, ...
2	ਪੈਟਰੋਲ ਦੀਆਂ ਕੀਮਤਾਂ ਨੂੰ ਮੱਦੇਨਜ਼ਰ ਰੱਖਦੇ ਹੋਏ \nਮੇ...	2021-02-24 09:23:22+00:00	0	0	https://twitter.com/parmarmaninder/status/1364...	[#FarmersProtest]
3	@ReallySwara @rohini_sgh watch full video here...	2021-02-24 09:23:16+00:00	0	0	https://twitter.com/anmoldhaliwal/status/13645...	[#farmersprotest, #NoFarmersNoFood]
4	#KisanEktaMorcha #FarmersProtest #NoFarmersNoF...	2021-02-24 09:23:10+00:00	0	0	https://twitter.com/KotiaPreet/status/13645061...	[#KisanEktaMorcha, #FarmersProtest, #NoFarmers...

Fig 1 Before tokenized content

Then we apply preprocess_document, a function that iterates over a dataset of tweets. For each tweet applies build_terms on its content.

	Tweet	Date	Hashtags	Likes	Retweets	Url
0	[world, progress, indian, polic, govt, still, ...	2021-02-24 09:23:35+00:00	[#ModiDontSellFarmers, #FarmersProtest, #FreeN...	0	0	https://twitter.com/ArjunSinghPanam/status/136...
1	[farmer, protest, modi, ignor, farmer, death, ...	2021-02-24 09:23:32+00:00	[#FarmersProtest, #ModiIgnoringFarmersDeaths, ...	0	0	https://twitter.com/PrdeepNain/status/13645062...
2	[ਆਰ, ਚ, farmer, protest]	2021-02-24 09:23:22+00:00	[#FarmersProtest]	0	0	https://twitter.com/parmarmaninder/status/1364...
3	[reallyswara, rohinisgh, watch, full, video, f...	2021-02-24 09:23:16+00:00	[#farmersprotest, #NoFarmersNoFood]	0	0	https://twitter.com/anmoldhaliwal/status/13645...
4	[kisan, ekta, morcha, farmer, protest, farmer,...	2021-02-24 09:23:10+00:00	[#KisanEktaMorcha, #FarmersProtest, #NoFarmers...	0	0	https://twitter.com/KotiaPreet/status/13645061...

Fig 2 Final output

2. Exploratory Data Analysis

Before doing the data analysis, we needed to construct a list of lists of terms with the content of all the pre-processed tweets.

2.1 Word counting distribution

After analyzing all the terms from the tweets, it is clear that the most frequently used terms in this collection are “farmer” and “protest,” which originate from the hashtag #FarmersProtest. This is expected, given that the dataset contains tweets related to the farmers' protest. We also observe that the other terms are significantly less frequent than these two, with none exceeding a value of 25000.

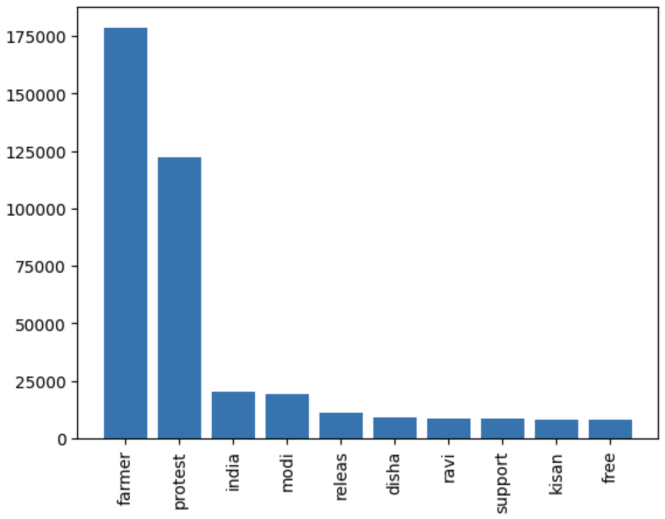

Fig 3 Graphic Word counting distribution

2.2 Average sentence length

The average sentence length of a tweet (counting the processed terms of a tweet) is 11.88 terms per tweet.

2.3 Vocabulary size

The vocabulary size of this collection of tweets is 57153 unique words.

2.4 Ranking of the most retweeted tweets

We can observe the top 10 retweeted tweets:

	Tweet	Retweets	Url
111329	मध्यप्रदेश में निजी व्यापारी 200 करोड़ का धान ...	7723	https://twitter.com/RakeshTikaitBKU/status/136...
7645	There's a #FarmersProtest happening in Germany...	6164	https://twitter.com/dhruv_rathee/status/136414...
89780	disha ravi, a 21-year-old climate activist, ha...	4673	https://twitter.com/rupikaur_/status/136088206...
88911	Disha Ravi broke down in court room and told j...	3742	https://twitter.com/amaanbali/status/136090860...
111556	Farmers are so sweet. Y'all have to see this @...	3332	https://twitter.com/jedijasmin_/status/1360162...
64492	india is targeting young women to silence diss...	3230	https://twitter.com/rupikaur_/status/136179092...
108072	Bollywood has betrayed Panjab & the farmer...	3182	https://twitter.com/RaviSinghKA/status/1360260...
60721	लहरों को ख़ामोश देख कर ये ना समझना कि समंदर मे...	3057	https://twitter.com/sherryontopp/status/136189...
29510	हाँ मैं जानता हूँ कि मैं शायर नहीं, और ज़ुल्म ...	3040	https://twitter.com/sherryontopp/status/136309...
24160	कलियुग है साहब , यहाँ झूठे को स्वीकार किया जा...	2622	https://twitter.com/sherryontopp/status/136337...

Fig 4 Top 10 retweeted tweet

2.5 Word clouds for the most frequent words

We can see the most repeated words clearly bigger. This is a visual representation of the most frequent terms in this tweet dataset. As we can observe given our tweet entries the most frequent words are 'farmer' and 'protest' and related words.Also we can spot 'india' related words too.

Fig 5 Word cloud of the most frequent words

2.6 Entity recognition

Named Entity Recognition (NER) is a crucial task in Natural Language Processing that involves identifying and classifying named entities in text into predefined categories such as names of persons, organizations, locations, etc. We created a simple NER system using the SpaCy library.

Once the pre-trained model is loaded and the named entity categories are specified, we use the model to identify and classify named entities in the text:

Fig 6 Entity recognition

Here we can see how our model identifies Govt as an organization, India as a Geopolitical Entity and Indian as a NORP, which stands for Nationalities or Religious/Political Groups.