NLP Based Analysis of Twitter Samples

Introduction

This report Analyses how NLP techniques are applied to a dataset of twitter's text analysis. The dataset contains positive, negative, and neutral tweets. The following are the objectives of this report :

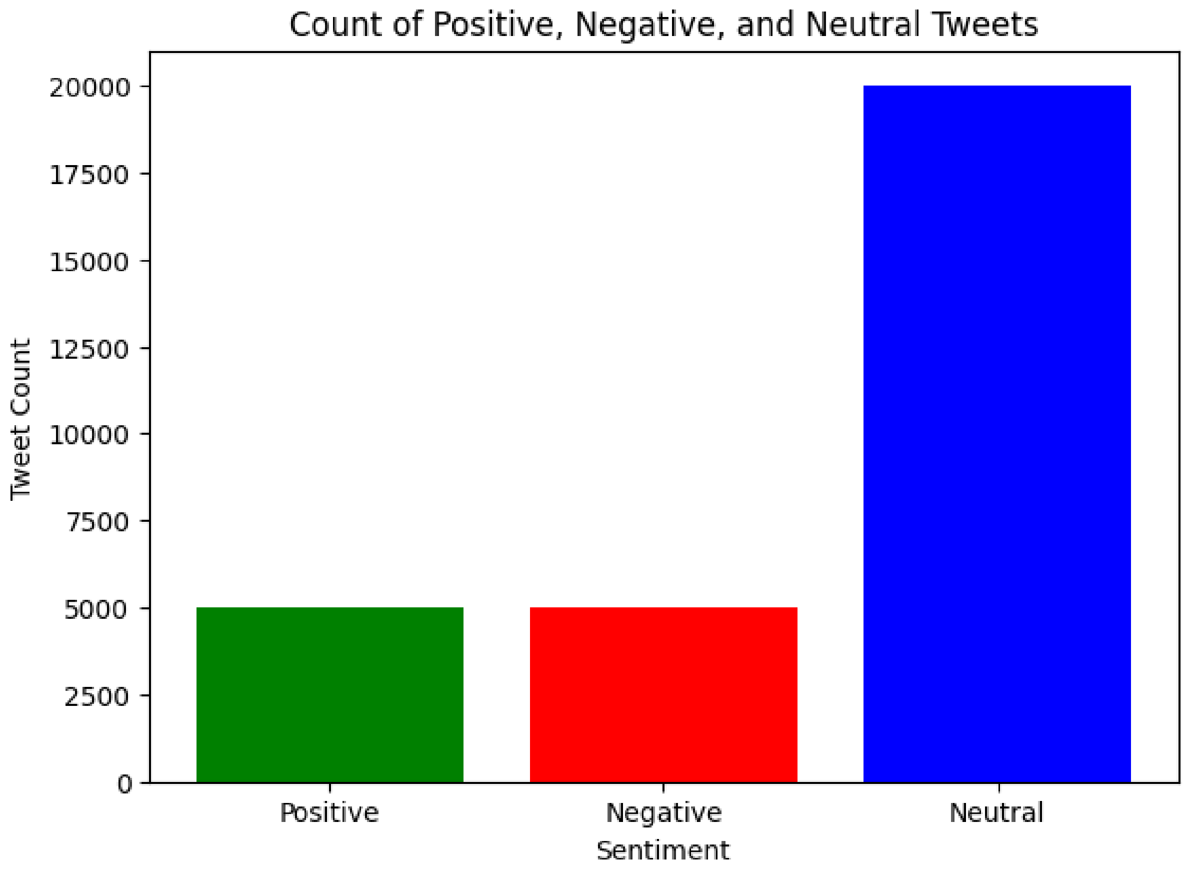
- Perform Exploratory Data Analysis (EDA)
- Preprocess the text data (Tokenization, Stopword Removal, Normalization, etc.)
- Build a Bag-of-Words model
- Analyze Part-of-Speech (POS) tagging
- Extract and analyze N-grams (unigrams, bigrams, trigrams)

These tasks performed are essential for preparing textual data for machine learning models and gaining insights into common patterns in the text.

1) Exploratory Data Analysis

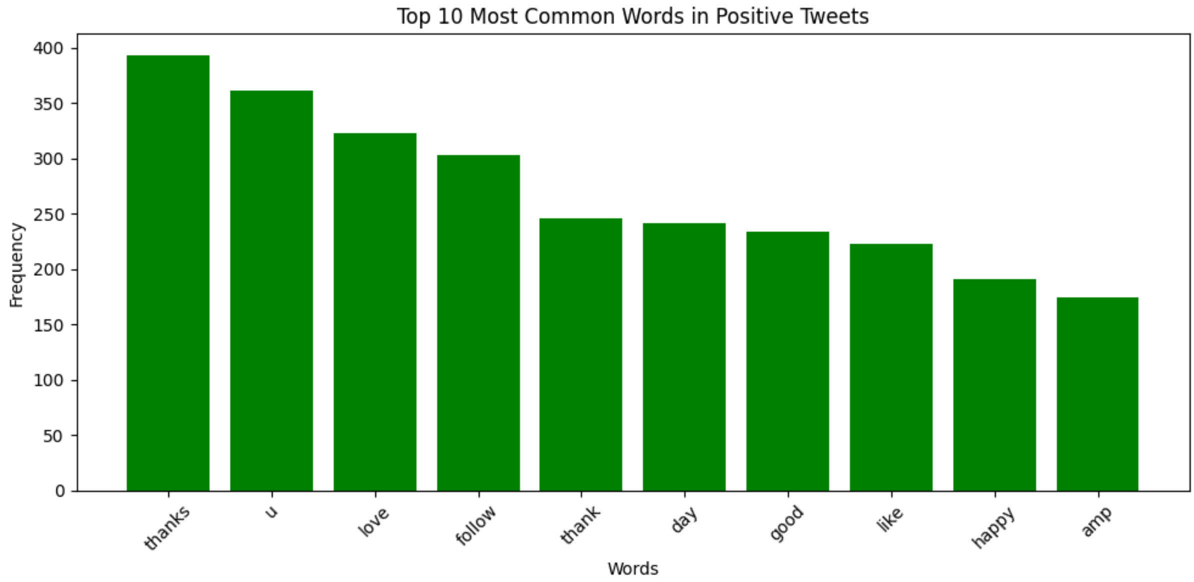
Exploratory Data Analysis (EDA) involves summarizing and visualizing the main characteristics of a dataset, improving the understanding trends in data before further analysis.

1.1) Counting the number of positive, negative, and neutral tweets.

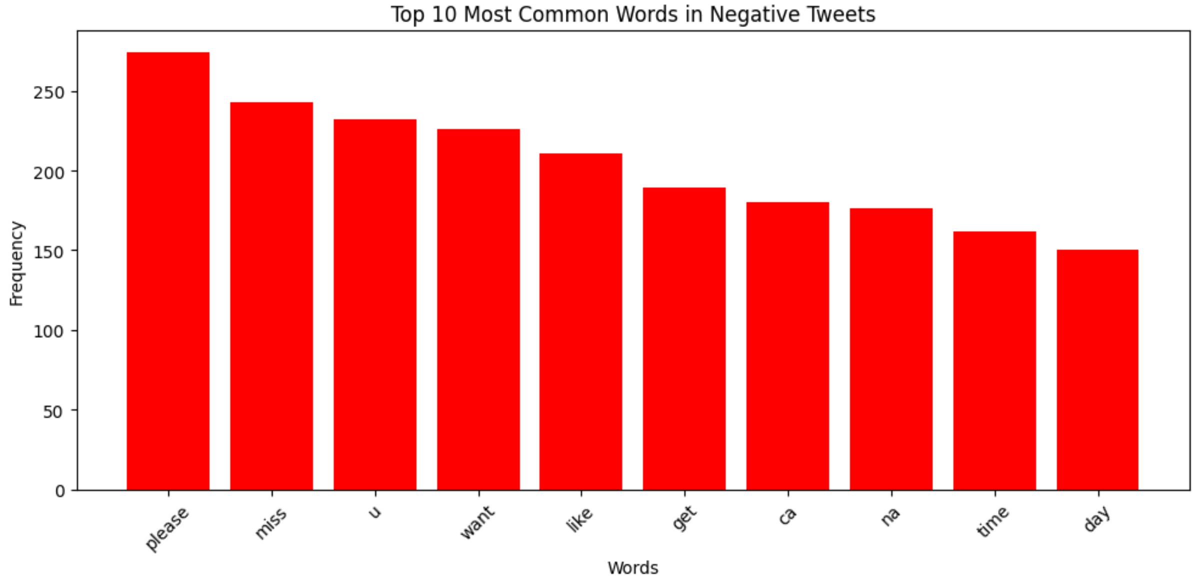


In the figure above neutral tweets have the most tweet count amongst others.

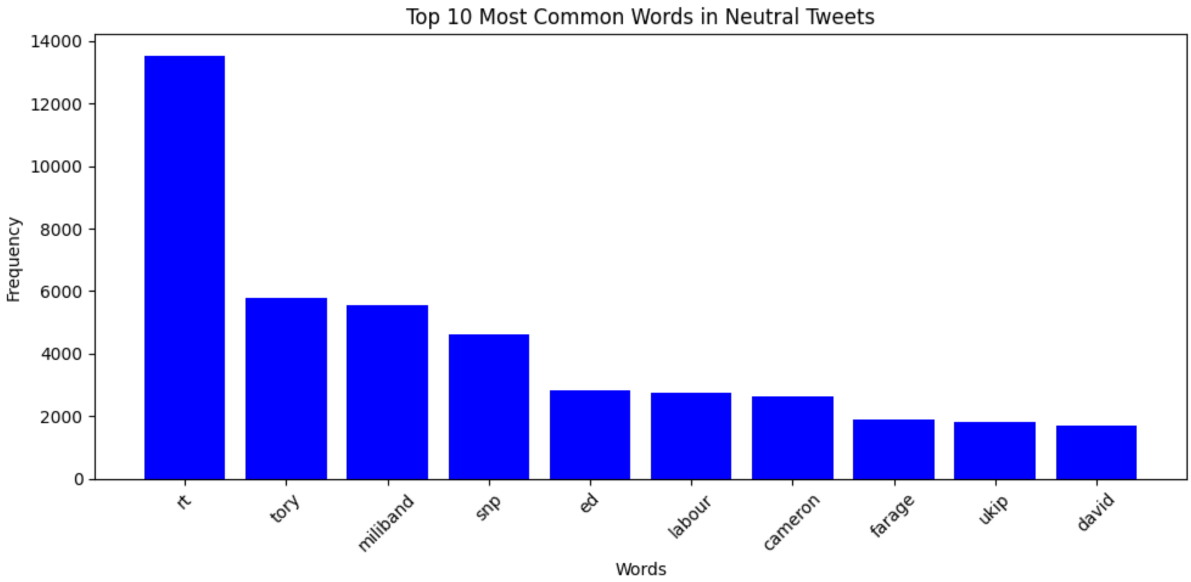
1.2) The frequency of words in positive, negative and neutral among Top 10 common words category



In positive tweets the most common word is “thanks” and the least is “amp” amongst the top 10.

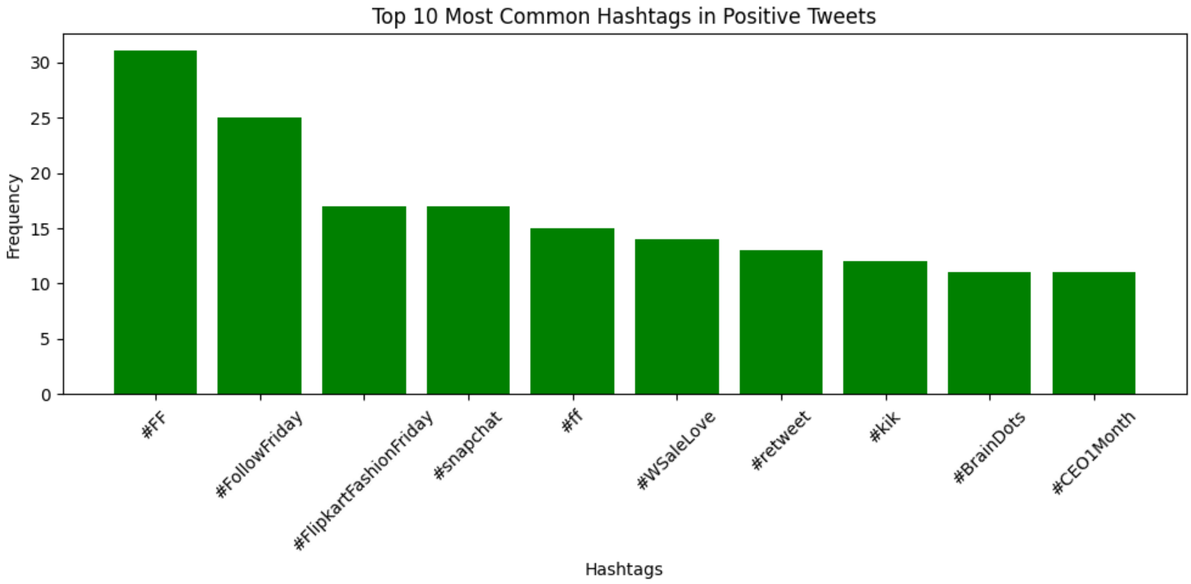


In negative tweets the most common word is “please” and the least is “day” amongst the top 10.

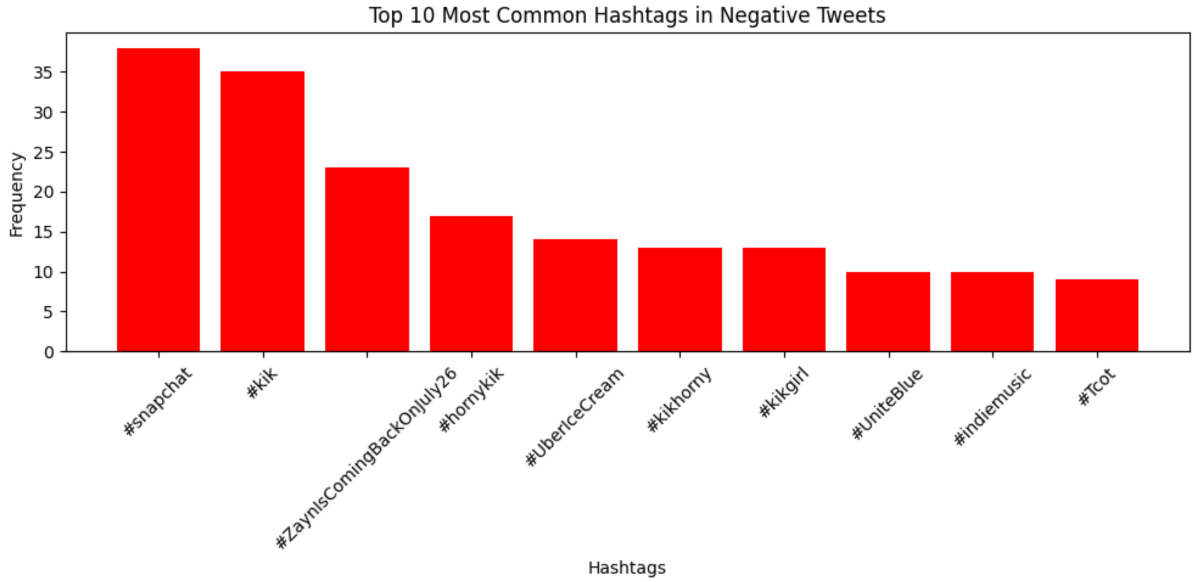


In neutral tweets the most common word is "rt" and the least is "david" amongst the top 10.

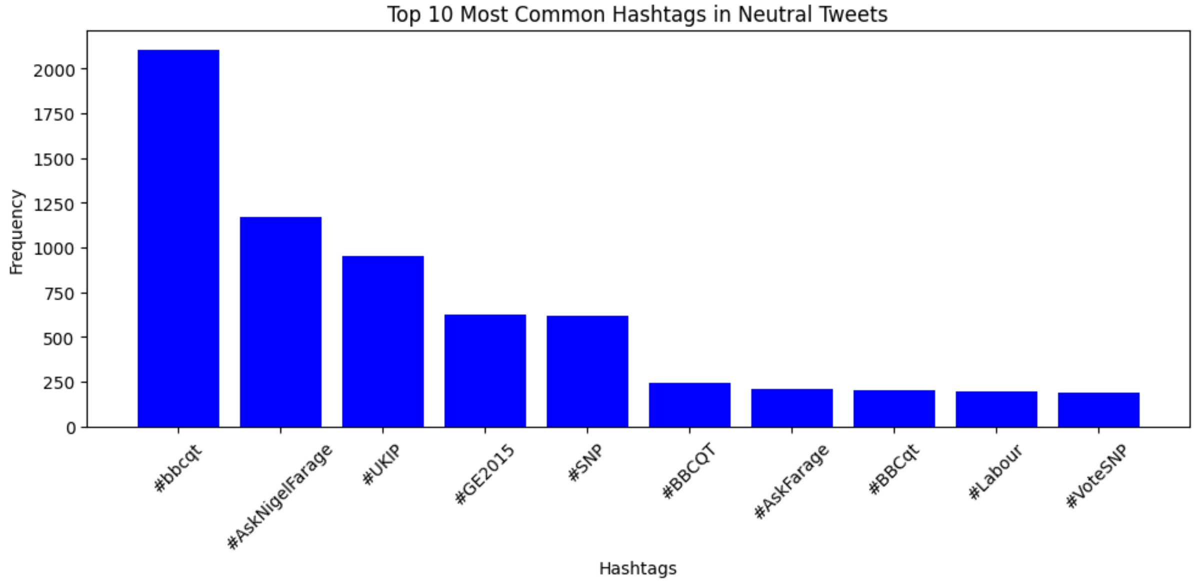
1.3) Frequency of hashtags in positive, negative and neutral tweets among Top 10 most common hashtags



In positive tweets the most common hashtag is “#FF” and the least is “#CEO1Month” amongst the top 10.

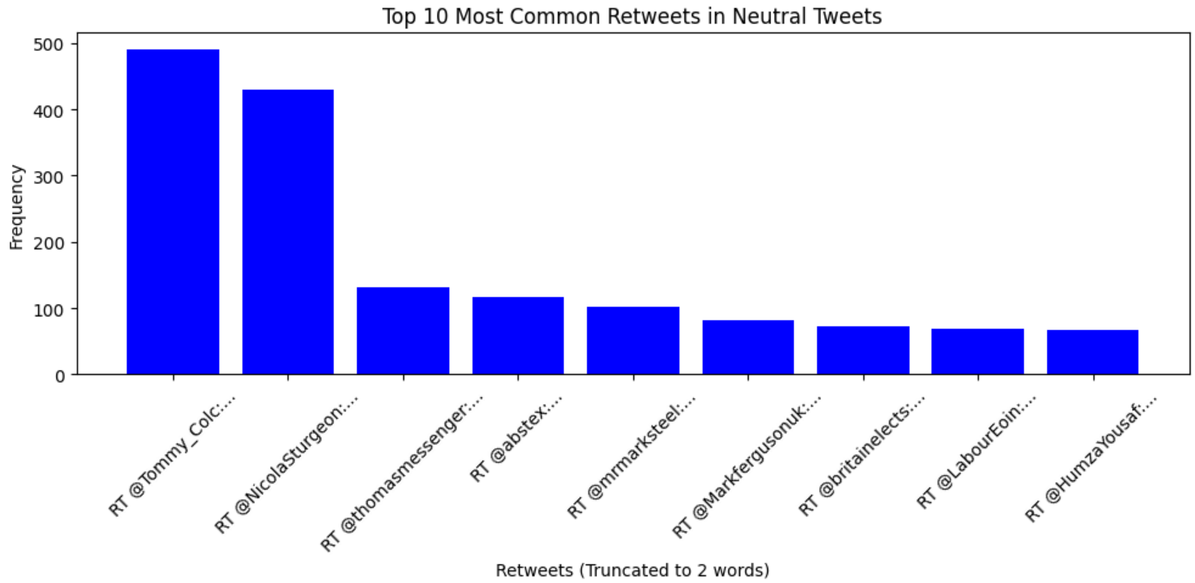


In negative tweets the most common hashtag is “#snapchat” and the least is “#Tcot” amongst the top 10.



In neutral tweets the most common hashtag is “#bbcqt” and the least is “#VoteSNP” amongst the top 10.

1.4) Frequency of retweets in positive, negative and neutral tweets among Top 10 most common retweets counts in each category



The most common retweets were only in the category of Neutral Tweets, and the one with the most frequency was “RT Tommy...” and the lowest one was “RT @HumzaYousaf..” in the category of top 10.

2) Preprocessing the test data

Preprocessing the text data involves cleaning and preparing raw text, that improve the quality of the text, making it ready for further analysis and modeling.

2.1) Tokenization:

A custom tokenizer was created to process tweets while preserving hashtags, mentions, URLs, and emoticons. A list of emoticons was defined and combined into a regex pattern, which was used to extract tokens from the tweets. The tokenizer was applied to a dataset, displaying the original tweets alongside their tokenized versions, ensuring important contextual elements are retained for analysis.

2.2) Stop words

A set of English stop words were loaded which helped to filter out from tokenized tokens through a function. The tokenizer output was iterated through, and for each tweet, the stop words were removed from the tokens. The cleaned tokens were then stored in a new list. Finally, the original tweets and their corresponding tokens with stop words removed were displayed, providing a clearer view of the meaningful content in the tweets.

2.3) Normalization

Normalization of tokens was done by converting them to lowercase. A new list was created to store the normalized tokens. A function was defined to process the tokens, converting each one to lowercase.

2.4) Extraction of words with Hashtag(#) and Mention(@)

A function was created to use regular expressions (RegEx) to identify hashtags (starting with #) and mentions (starting with @). The original text, including these elements, was stored alongside the extracted data to verify and furthermore these results were compiled into a new list and displayed highlighting the relevant social media interactions.

2.5) Extraction of URLS, and replace with URLS

In this extraction URLs from tweets are replaced with the token "url." A function is created using regular expressions (RegEx) to identify URLs starting with http or https and capture them.

2.6) Emoticon and Emoji Removal

The removal of emoticons and emojis from tweets is done to enhance text quality for analysis. It uses a regular expression pattern to identify common emoticons and defines a function, remove_emoticons, to clean the text.

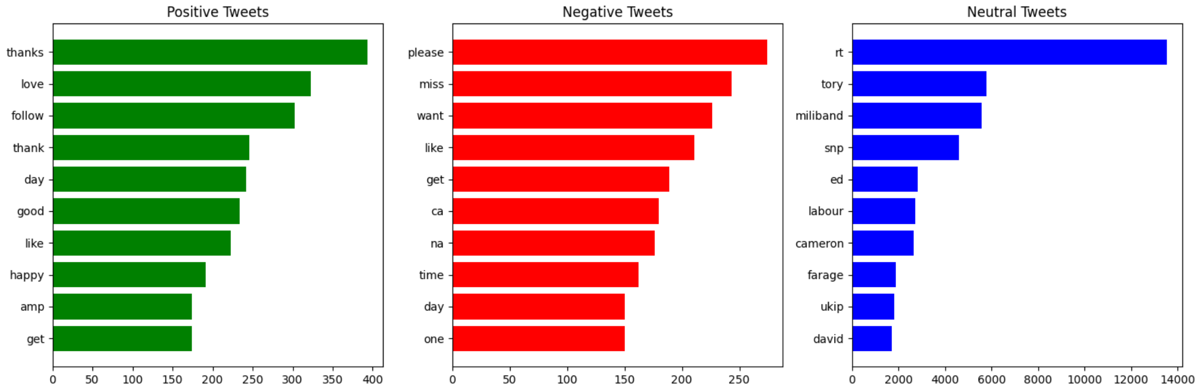
3) Bag of words (BOW) representation on processed text data

Bag of words transforms text into a numerical format for analysis. It converts documents into vectors based on the frequency of each word, ignoring grammar and word order.

3.1) BOW to identify positive, negative, and neutral tweets using CountVectorizer

In this section the pre-processed tweets were combined for a Bag-of-Words (BoW) representation, which was created using CountVectorizer and displayed as a DataFrame for better visualization.

3.2) Graphical representation of top 10 in each category



In top 10 positive tweets the highest frequency was “thanks”, and the least was “get”.

In top 10 negative tweets the highest frequency was “please”, and the least was “one”.

In top 10 neutral tweets the highest frequency was “rt”, and the least was “david”.

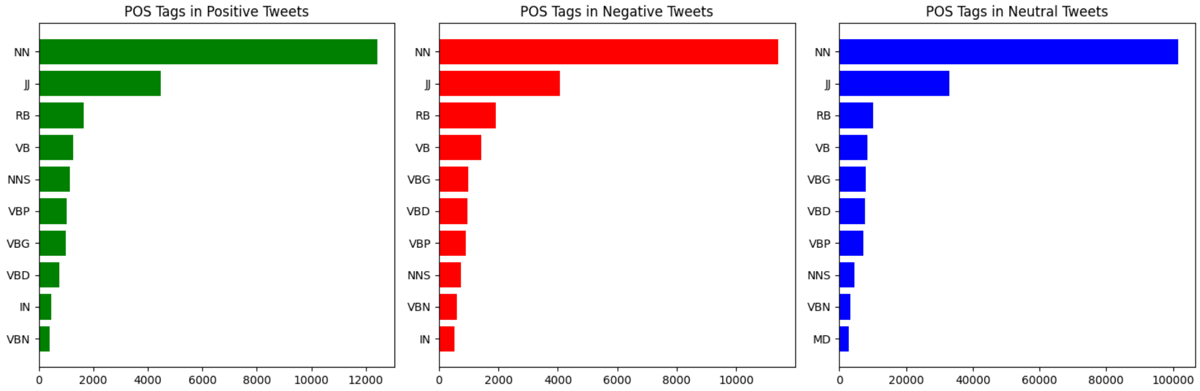
4) Perform part of speech (POS) tagging on the preprocessed text data

Part-of-Speech (POS) tagging assigns labels to each word in a text based on its grammatical role, such as noun, verb, adjective, or adverb. This tagging helps to identify the structure of sentences, facilitating better understanding of the text’s meaning and context.

4.1) Most common POS tags in positive, negative, and neutral tweets at words level

In this section, tweets were preprocessed to include Part-of-Speech (POS) tagging. POS tagging was applied to the filtered words. The most common POS tags were counted and extracted into dictionaries providing insights into the linguistic structure of the tweets across different sentiments.

4.2) Graphical Visualization of top 10 in each category



In top 10 positive POS Tags the highest frequency was “NN”, and the least was “VBN”.

In top 10 Negative POS Tags the highest frequency was “NN”, and the least was “IN”.

In top 10 Negative POS Tags the highest frequency was “NN”, and the least was “MD”.

So we can say “NN” has high frequency in all of the categories.

5) N-grams to identify common phrases in positive, negative, and neutral tweets

N-grams are contiguous sequences of N items—typically words or characters—from a given text. By utilizing N-grams, such as bigrams (2-grams) and trigrams (3-grams), It captures common phrases and contextual relationships within the tweets, which single words may not convey effectively. It analysis allows for a deeper understanding of the linguistic patterns in tweets.

5.1) Lemmatization to reduce words to their base form at pre-processed text

Lemmatization was applied to the preprocessed tweets using the WordNet lemmatizer. The apply_lemmatization function reduces words to their base forms, normalizing variations for improved analysis. Lemmatization was performed on positive, negative, and neutral tweets, and the resulting lemmatized words were printed

5.2) N-grams from preprocessed text data

N-grams (unigrams, bigrams, and trigrams) from the lemmatized tweets. Unigrams were directly taken from the lemmatized words, while the generate_ngrams function was used to create bigrams and trigrams by combining adjacent words. Unigrams were generated for positive, negative, and neutral tweets without any additional processing. Bigrams and trigrams were generated separately for each category, capturing word pairs and triplets, respectively.

5.3) Frequency of each N-gram

The frequencies of N-grams (unigrams, bigrams, and trigrams) from the lemmatized tweets was calculated using calculate_ngram_frequencies function. This function initializes a counter to track the frequency of each N-gram as it processes the tweets.

Unigram frequencies were calculated for positive, negative, and neutral tweets and same process was applied to obtain bigram and trigram frequencies for each sentiment category.

5.4) N-grams sorted in descending order

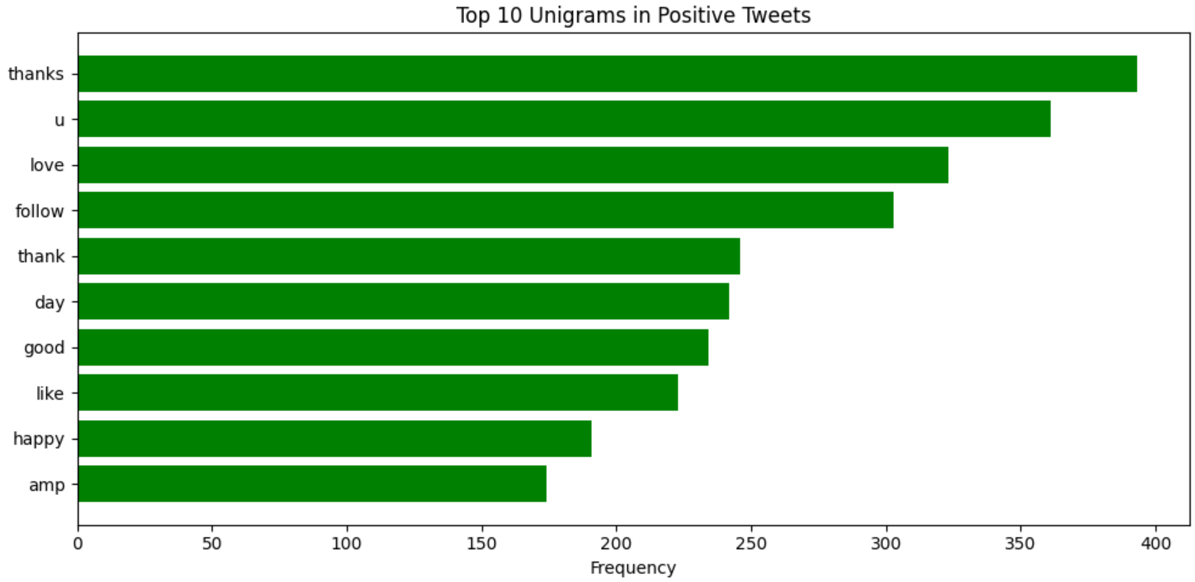
The N-grams (unigrams, bigrams, and trigrams) were sorted by frequency using the most_common() method. This method allows us to retrieve the N-grams in descending order of their occurrence, providing a clear view of the most frequently used terms in the tweets.

5.5) Top 10 most frequent N-grams for each type of n-gram

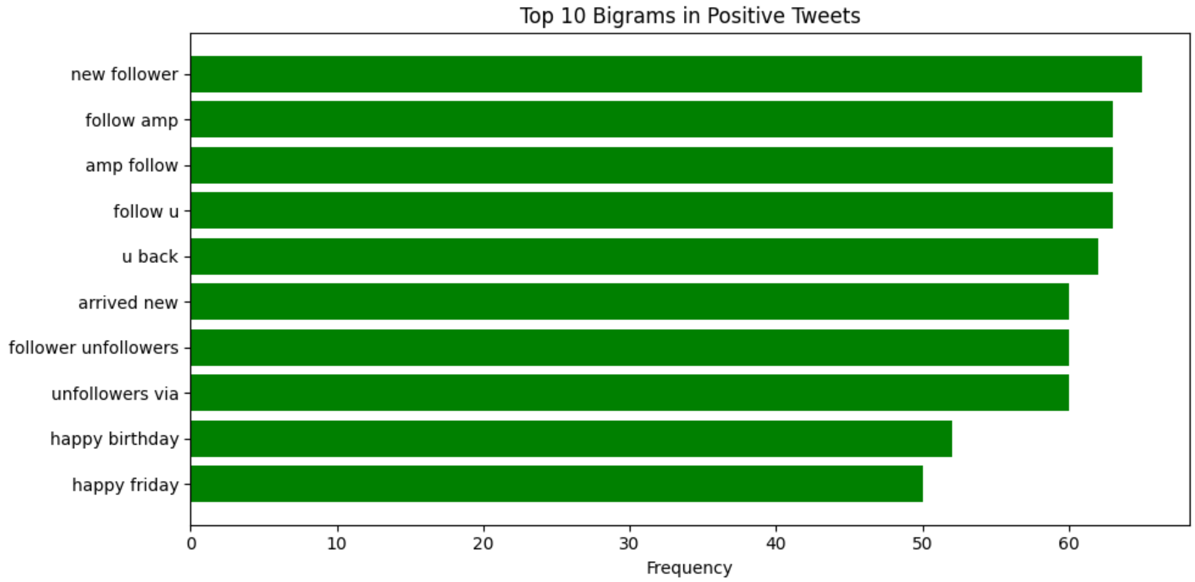
The top 10 unigrams were extracted for each category, followed by the same process for bigrams and trigrams.

5.6) Graphical of top 10 N grams in each category:

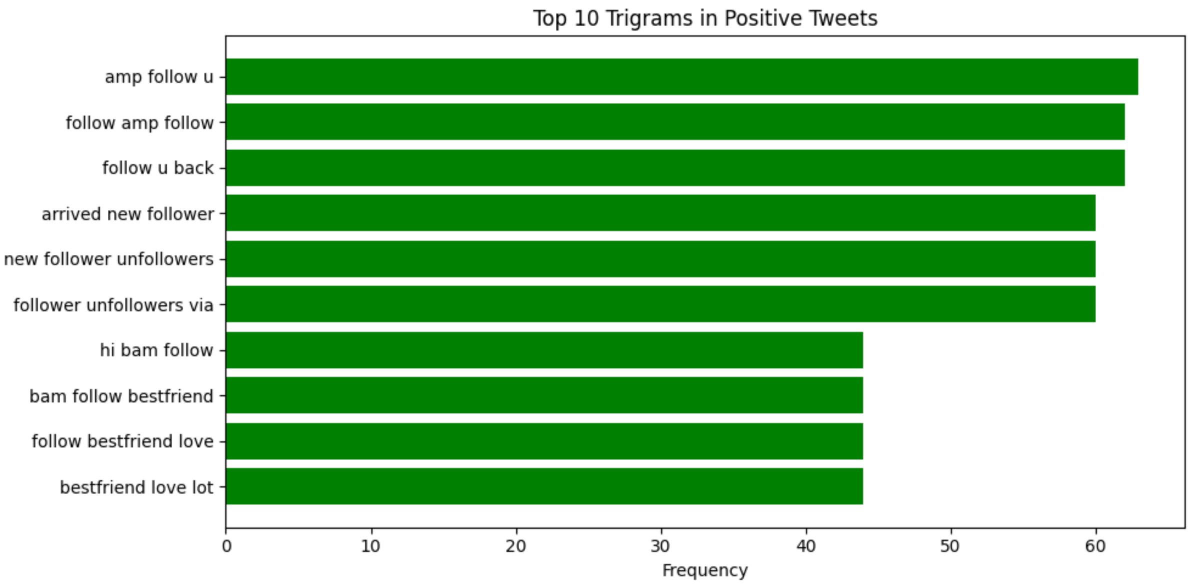
1) For the case of Positive Tweets:



In top 10 unigram the highest frequency was of “thanks”, and the least was of “amp”.



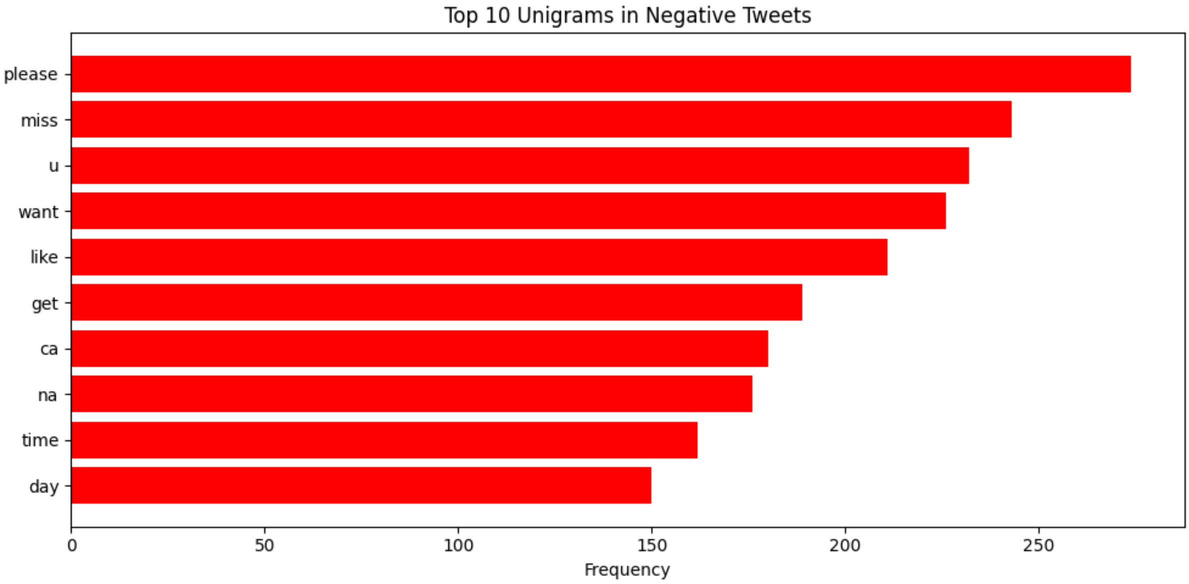
In top 10 bigram the highest frequency was of “new follower”, and the least was of “happy Friday”.



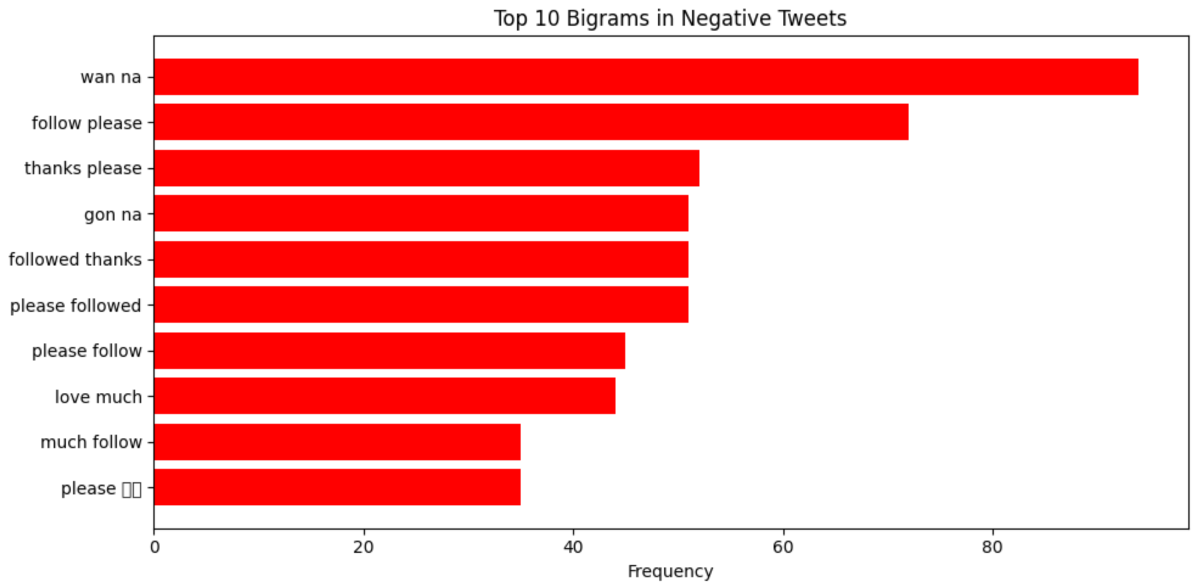
In top 10 trigram the highest frequency was of “thanks”, and the least was of “bestfriend love lot”.

And just like that there is a comparison with the category of negative and neutral.

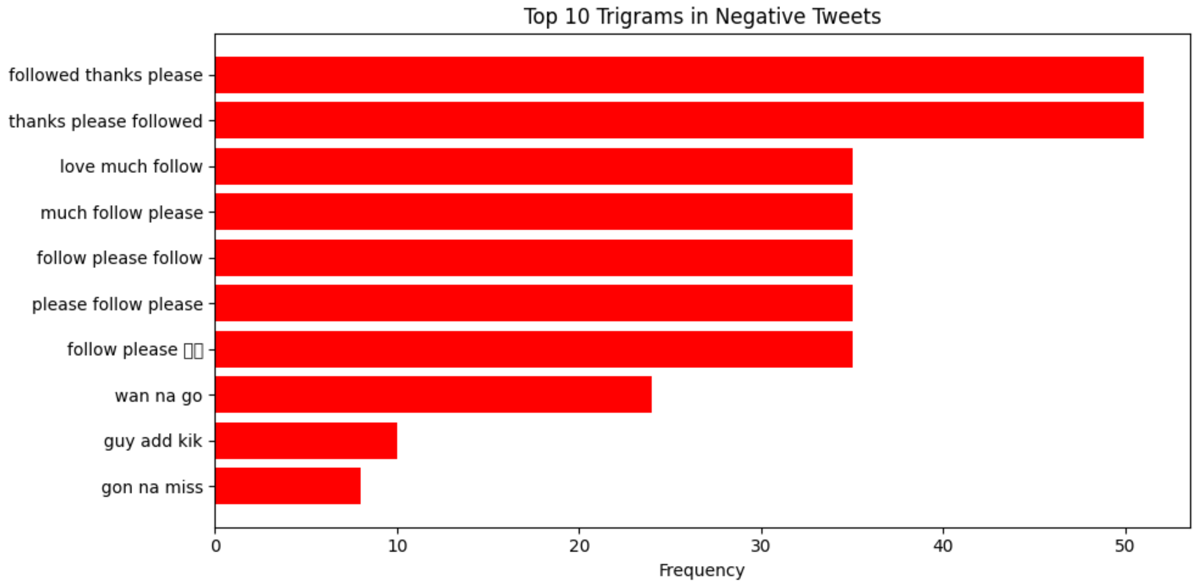
2) For the case of Negative Tweets:



In top 10 unigram the highest frequency was of “please”, and the least was of “day”.

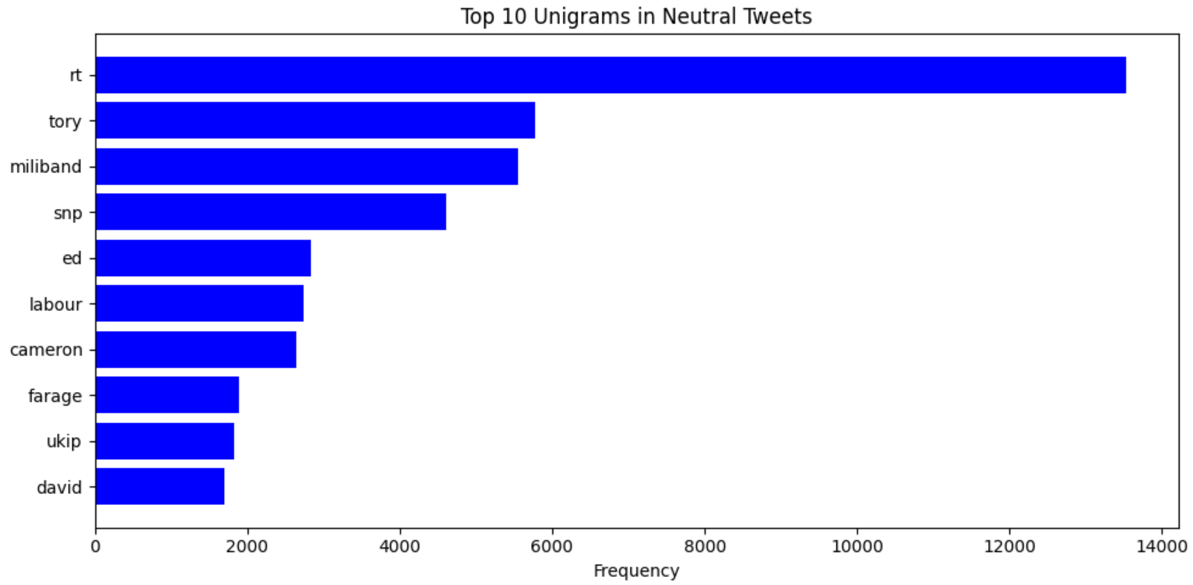


In top 10 bigram the highest frequency was of “please”, and the least was of “day”.

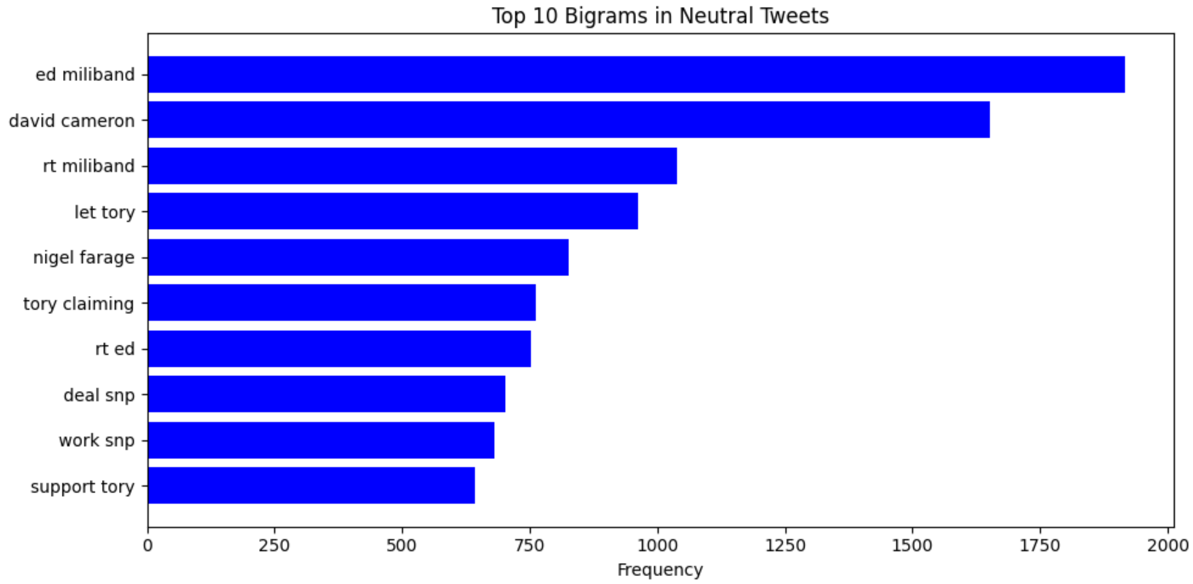


In top 10 trigram the highest frequency was of “followed thanks please”, and the least was of “gon na miss”.

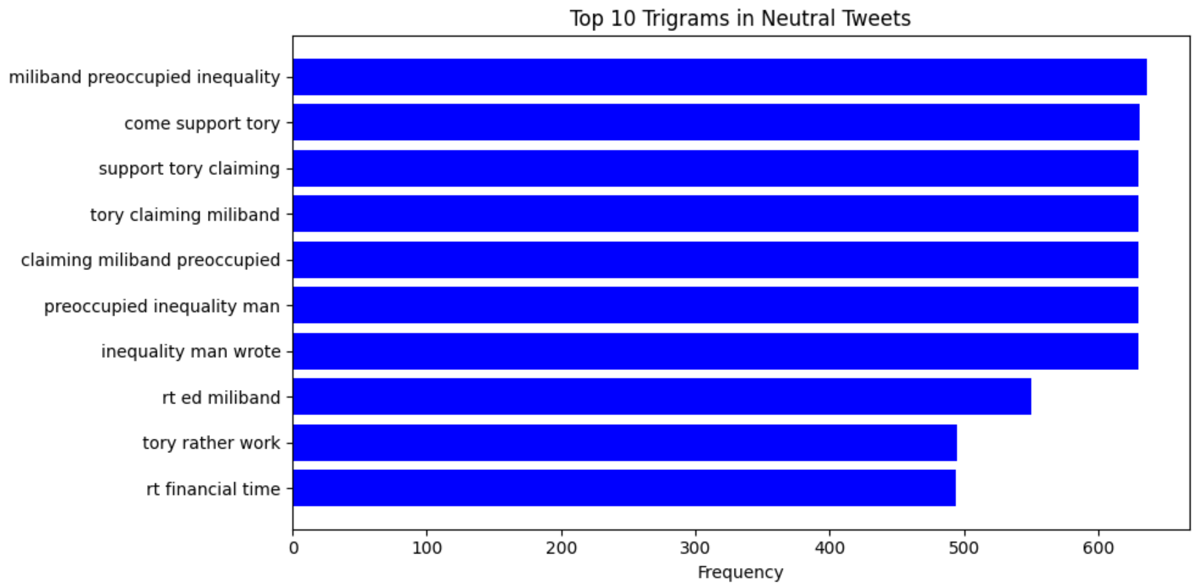
3) For Neutral Tweets:



In top 10 unigram the highest frequency was of “rt”, and the least was of “david”.



In top 10 bigram the highest frequency was of “please”, and the least was of “support tory”.



In top 10 trigram the highest frequency was of "milband preoccupied inequality", and the least was of "rt financial time".

Conclusion

In conclusion, this report presents a thorough analysis of Twitter data through Natural Language Processing (NLP) techniques, focusing on sentiment classification of tweets. The exploratory data analysis (EDA) provided insights into the distribution and characteristics of various sentiments. Key preprocessing steps, were implemented to enhance the quality of the text data. The Bag-of-Words model was employed to represent the processed tweets numerically, facilitating further analysis. Additionally, Part-of-Speech (POS) tagging and N-gram analysis were conducted to uncover linguistic patterns and common phrases within the tweets. Overall, the analysis demonstrates the effectiveness of NLP methods in extracting meaningful insights from social media data, contributing to a deeper understanding of sentiment-driven language use.