Syracuse University

CIS 600 - Applied Natural Language Processing
Spring 2024

Term Project Report

Sentiment Analysis On Syrian Civil War Using Reddit Data

Team Members :

Venkata Sri Siva Ramakrishna Palaparthy - 433976193
Nikhilesh Gunnam - 502430168
Prathik Reddy Sannapureddy - 344477369
Girish Kanjiyani - 587443048
Ruiying Chen - 255618384

Under the guidance of :
Prof. Edmund Yu

SYRACUSE UNIVERSITY · SUOS CULTORES SCIENTIA CORONAT · FOUNDED A·D· 1870

SYRACUSE UNIVERSITY ENGINEERING & COMPUTER SCIENCE

CONTENTS

INTRODUCTION

This project aims to do a comprehensive sentiment analysis of discussions about the Syrian conflict using data collected from Reddit. Finding out what Reddit users thought of this bitter and ongoing debate is the main objective. Sentiment analysis was employed in this study to bring out the influence of real-world events on social media platforms and to locate the attitudes, opinions, and feelings which were prevalent in these discussions. The thoughts concerning the Syrian war that circulated on Reddit were extensively examined in this study using sentiment analysis techniques. The practice of meticulously identifying and categorizing the attitudes and feelings included in textual data is known as sentiment analysis. To determine if Reddit users mostly expressed neutral, positive, or negative sentiments, this study employed computational tools and procedures. By applying analytical tools and approaches, this study aimed to ascertain if those who use Reddit mostly expressed beneficial, bad, or neutral opinions regarding various aspects of the war.

The Syrian war is a major global geopolitical issue with multiple dimensions, especially political, historical, and humanitarian ones. This experiment's specific objective was to find out how individuals on Reddit and other social media platforms view and discuss this conflict. The speeches often reflect a range of perspectives, personal experiences, biases, and emotions around the topic at hand. An understanding of the wide variety of viewpoints and emotions that participants in these discussions must contribute can be gained by looking at these sensations. The study's results were later expressed through some analyses and visualizations. Tools like word clouds, pie charts, bar graphs, and sentiment growth graphs were used to display the spread and trends of sentiments across the Reddit data set. The graphical representations helped to clarify the most common emotions, patterns of emotional activity, and likely changes in mood over time in the larger context of Reddit conversations regarding the Syrian war.

In conclusion, I wanted to show how actual events, like the violence in Syria, are related to how they are seen on social media. The goal was to show how these events appear, influence, and shape conversations on the internet through a sentiment analysis of Reddit. To show how these events appear, influence, and shape the online conversation, a study of Reddit sentiments was conducted. Understanding these sensations helps comprehend how social media platforms are impacted by real-world events since they reflect the diverse perspectives, emotions, and reactions found in the online discourse around this controversy.

METHODOLOGY

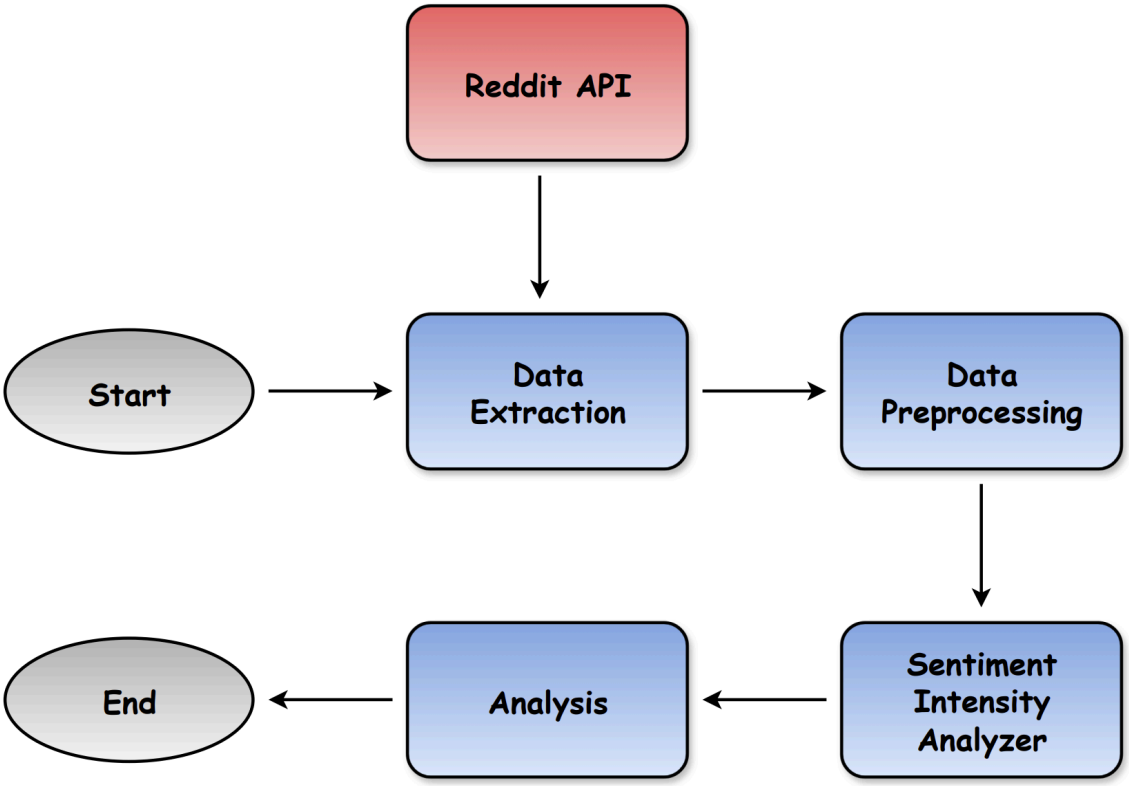
To examine Reddit users' opinions and the conversations around the situation in Syria, our team employed the following methodology :

- **Data Extraction :** By using this Reddit API, we were able to see Reddit conversations on the Syrian crisis by concentrating on forums or threads. Our Python app collected text-based data, timestamps, and user information by Reddit's API specifications.

- **Data Preprocessing :** After the extraction process, the data was cleaned to get rid of noise, duplicates, and useless posts. Text standards were applied, including tokens for quick evaluation and particular character deletion.

- **Sentiment Intensity Analyzer :** Applying sentiment analysis techniques like VADER, we assessed what was being said and expressed in every text section gathered from Reddit discussions. These methods generated neutrality ratings (positive, negative, and neutral) according to the sentiment and word context.

- **Analysis :** The post-sentiment rating data showed the difference between positive, negative, and neutral thoughts. The use of word clouds and bar graphs are the main visual methods used for demonstrating the overall opinion pattern and offering data on common sentiments of Reddit users discussing the Syria issue.



DATA COLLECTION

WHY REDDIT ?

Reddit is an online media and chat platform where users comment on articles and edit them together. The website's name is a play on the phrase "I read it." (TechTarget). On this platform, there are many subreddit communities. Every Reddit subreddit focuses on one particular topic, for example, music, politics, or technology. The most frequently read articles in every normal subreddit make up Reddit's front page or just the front page as it is sometimes referred to. The preset standard list includes the following subreddits: "pics," "funny," "videos," "news," and "gaming." Reddit's style makes analysis of data and access easier than on other social media sites. Researchers can get data on postings, debate threads that have active user demographics, and engagement metrics using the system's APIs. On top of that, Reddit's layout makes it easier to find famous people, or "power users," with a large following and may impact other people's thoughts and actions. PRAW, a Python shell for the Reddit API, lets users gather information from subreddits on Reddit.

With over 100,000 active topic groups, or "subreddits," and over 50 million daily active members, Reddit is one of the most popular social networking sites. A growing number of studies over the last ten years have used Reddit as a data source due to its popularity, accessibility, and capacity to provide high-quality data. Among the numerous resources were the original article, discussion threads and comments, meta-data, media, suggestions for upvoting and downvoting, subreddit features, and surveys of users and moderators.

DATA COLLECTION FROM REDDIT

The Python Reddit API Wrapper, or PRAW for short, is a Python module that lets users see Reddit feedback, discussions, and other data. Programmers can use this library to communicate with the Reddit API more easily by obtaining, examining, and adding to the huge quantity of data shared among the Reddit community. Researchers and developers can quickly and freely obtain Python content from Reddit with the aid of the API. This material can be merged into bigger collections of data.

Developers can communicate with Reddit's extensive database programmatically thanks to a set of guidelines and tools called the Reddit API (Application Programming Interface). It offers an organized method of accessing all of the platform's features, including posts, comments, user data, and more. By acting as a conduit among developers with this API, PRAW makes it easier to include Reddit's features into Python apps.

```
reddit = praw.Reddit(
    client_id='your_client_id',
    client_secret='your_client_secret',
    user_agent='your_user_agent'
)

subreddit = reddit.subreddit('your_subreddit')
```

Developers must use the Reddit Developer Portal to get an active Reddit app before beginning PRAW. Obtaining the user ID and user secret is required in this phase in order to create an encrypted link among the application and Reddit's servers for authentication. Developers can use the required login settings to launch a PRAW session whereas the credentials are specified. The Reddit example facilitates access to the various capabilities provided by the Reddit API. Following login, developers can play around with different features to customize apps to users' requirements.

Reddit's content is organized into posts, comments, and subreddits. Private online communities called subreddits are created with a specific purpose in mind. Contributions from posts and comments support the continuing discussions in these groups. PRAW is able to recognize data at several levels of this system, making it easier to go through and simplifying the process. The information displayed below was taken from the Syria Community subreddit.

DATA SOURCES FOR OUR RESEARCH

Sentiment analysis regarding the Syrian war utilizing data from relevant Reddit boards forms the basis of our approach. The idea is simple: use the Reddit API to generate CSV files (data files) including conversation, vote, and comment data, and then establish a couple of groups on Reddit as a data source.

```
subreddit = reddit.subreddit("AskMiddleEast")
limit_per_request = 1000
total_limit = 10000
all_posts = []

# Fetch posts in chunks until the desired total limit is reached
while len(all_posts) < total_limit:
    # Calculate the remaining limit for the current request
    remaining_limit = total_limit - len(all_posts)

    # Adjust the limit for the current request based on the remaining limit
    current_limit = min(limit_per_request, remaining_limit)

    # Fetch the top posts using the current limit
    top_posts = subreddit.new(limit=current_limit)

    # Extend the list of all_posts with the current set of posts
    all_posts.extend(top_posts)

    # Print a message to indicate progress (optional)
    print(f"Fetched {len(all_posts)} out of {total_limit} posts")
```

```
posts_data = []
count = 0
for post in all_posts:
    post_info = {

        "title": post.title,
        "score": post.score,
        "id": post.id,
        "url": post.url,
        "comms_num": post.num_comments,
        "created": post.created_utc,
        "body": post.selftext,
        "timestamp": post.created_utc,
        "subreddit" : post.subreddit.display_name
    }
    posts_data.append(post_info)
    count += 1
df = pd.DataFrame(posts_data)
print(df)
```

The necessary data can be acquired from Python objects in the following manner: it is first fetched into lists, which are subsequently taken into data frames, and last it is converted to CSV.

The steps followed for the Data Extraction from reddit is as follows :

- Importing Libraries like PRAW, pandas
- Setting Up Reddit API Credentials
- Accessing Subreddit - “Syria”, “SyriaWar_23” etc.
- Setting up parameters such as limit and number of requests
- Fetching posts in chunks
- Collecting post information in the form of python objects
- Create a dataframe
- Convert the dataframe into csv and csv names are assigned accordingly.
- Merge data into dataframe before data preprocessing.

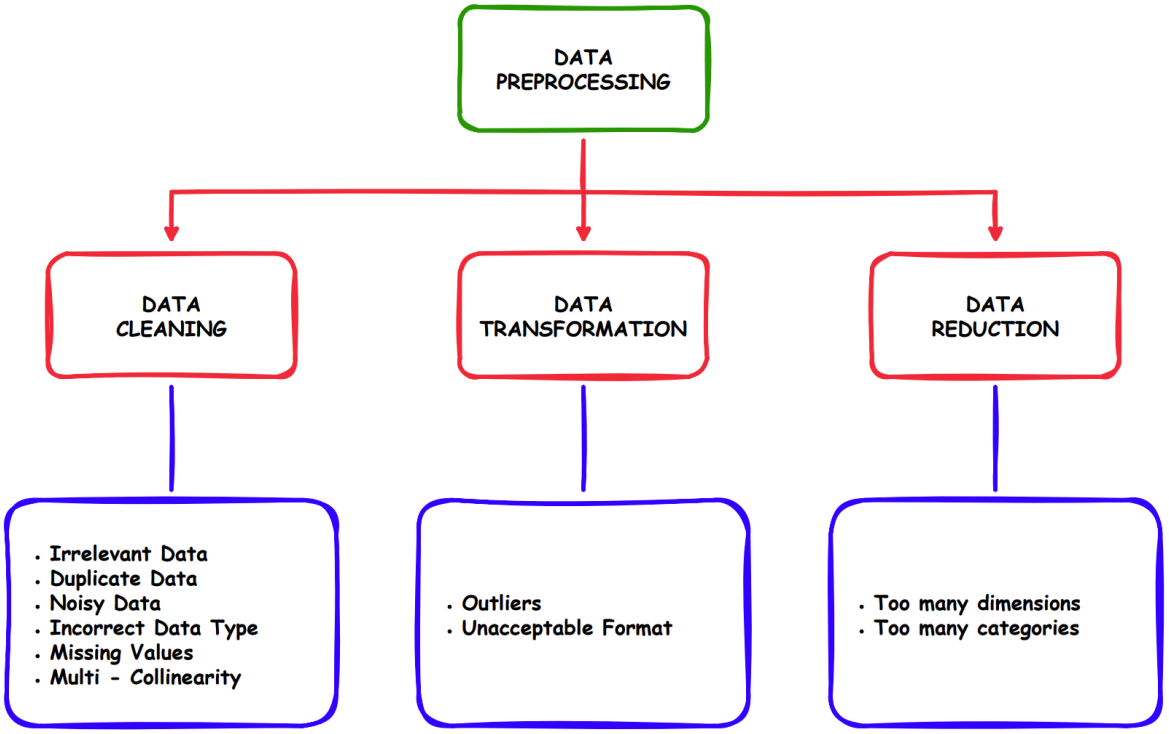
DATA PREPROCESSING

DATASETS AFTER MERGING

```
merged_df.sample(5)
```

	title	score	id	url	comms_num	created	body	timestamp	subreddit
16358	The curse strikes again	175	sbyrz0	https://i.redd.it/pjaytobvrpd81.png	3	1.643064e+09	NaN	2022-01-24 22:43:10	SyrianCirclejerkWar
30200	A new blast leaves more than 25 people dead or...	1	eeoj71	http://www.syriahr.com/en/?p=151248	0	1.577126e+09	NaN	2019-12-23 18:36:59	syrianconflict
409	Are East Asian men seen as cute in the Middle ...	3	1brq1eb	https://www.reddit.com/r/AskMiddleEast/comment...	13	1.711827e+09	Because we have baby face and no beard, do you...	2024-03-30 19:26:57	AskMiddleEast
11828	Turkish Space Agency sponsored by Mountain ker...	58	pajnu8	https://i.redd.it/7l3o7hu8v9j71.jpg	0	1.629796e+09	NaN	2021-08-24 09:09:58	SyrianCirclejerkWar
17248	"Just minor lossed comrade"	10	t6gpya	https://streamable.com/16l49r	3	1.646392e+09	NaN	2022-03-04 11:14:40	SyrianCirclejerkWar

For every machine learning project or study, data preparation is essential to ensuring dependable, consistent, and easily analyzed data.The process of cleaning, converting, and arranging unprocessed data into a structure that can be used for analysis is known as information preparation. Mistakes with raw data, missing figures, and inconsistent results can complicate research using our methods. Preprocessing adds missing data, corrects formatting errors, and eliminates extraneous information to help identify and address these problems. Improving the validity and quality of the data makes analysis and information collecting easier. A distinct type of material preparation, known as feature scaling, is usually produced from the data and involves classifying numerical features into a single category in order to minimize bias in the model's output.



To ensure that the data is reliable, consistent, and expressed in a format that is easy to understand, gathering the information is thus necessary. By doing data preparation, you can make sure that your own research or machine learning models are founded on accurate information that can generate dependable views and predictions.

The sentiment study of Reddit API conversations on the Syrian war gives an array of data. However, sometimes noisy, incomplete, or unused information obtained from Reddit may affect the validity and precision of sentiment analysis outcomes. In this case, preprocessing the information is very essential to deal with issues like crossover and outliers, manage missing data, and remove unnecessary or noisy data. Duplicate Data: Duplicate data should be removed even if the proportion of sentiments expressed in discussions may remain the same because it can distort the analysis's conclusions and burden computers more.

One of the project's preprocessing steps for sentiment analysis was copying relevant columns. removing NaN values because they reinforce the idea of zero Missing data might occur in the absence of posts or comments. Because missing data can skew the results of the study, it should be addressed by either removing the affected data or inserting the missing values using interpolation or other approaches. information converting all text to lowercase to preserve consistency and removing handles that start with "@" since they are only mentions and don't help with sentiment analysis removing URLs (which are only for reference) and applying a few other processing methods, such removing single and special characters, swapping out multiple spaces for a single space, and formatting timestamps in datetime. By effectively cleaning and structuring the information, removing unnecessary or noisy data, resolving data that is lacking, addressing crossover and outliers, and reducing the number of variables and groups, we can ensure the value and reliability of sentiment analysis results.

SENTIMENT ANALYSIS AND FEATURE ENGINEERING

One way to assess how well an emotion is portrayed in each text is to look at the sentiment intensity of the assessment itself. This natural language processing (NLP) method assists in identifying the attitude—whether neutral, positive, or negative—of a written piece.

Sentiment intensity analysis is critical to many applications, including market research, online conversation monitoring, image management, and customer input evaluation. Sentiment intensity analysis looks at a sentiment's psychological significance in order to present a more complex picture of an emotion than just a good, negative, or neutral one.

An algorithm called the Sentiment Intensity Analyzer assigns a text a score based on how strongly a specific sentiment is expressed in it. On the scale, 0 represents a neutral feeling, +1 indicates a strongly positive sentiment, and -1 indicates an intensely negative sentiment.

To ascertain sentiment intensity, the Sentiment Intensity Analyzer frequently blends rule-based and machine learning approaches. To ascertain the overall sentiment strength of a given text, it considers a number of textual elements, including word order, capitalization, punctuation, and the existence of positive or negative expressions.

Sentiment intensity analysis's primary benefit is that it provides a deeper understanding of the sentiment expressed in a text. A sentence that contains both positive and negative terms could be given a sentiment intensity rating instead of being clearly classified as positive or negative. The text's overall emotional impact is reflected in this score.

Ratings of sentiment intensity are useful for monitoring changes in consumers' opinions about a good or service. Sentiment intensity ratings can be used to monitor how consumers' perceptions of a specific brand or product vary over time, as well as how their sentiments alter. The Sentiment Intensity Analyzer is one such instrument. Sentiment analysis is a helpful technique for determining the emotional content of the text.

WHY IS SENTIMENT ANALYSIS DIFFICULT TO PERFORM ?

Even though sentiment analysis seems simple on paper, it's a complex topic. Texts can convey more than one emotion at once. As an example,

“The acting was good , but the movie could have been better”

The above sentence consists of two polarities. In our case, the main topic/word “Ethereum” is used in both the polarities.

VADER SENTIMENT ANALYSIS

Valence Aware Dictionary for Sentiment Reasoning (VADER) is a well known sentiment analysis model for dealing text data. This model can detect emotional polarity which are either positive or negative, and intensity i.e. strength of the emotion. The VADER package is available in the NLTK package. This could be used directly for any unlabeled textual data. VADER sentiment analysis makes use of the lexicon in order to integrate sentiment scores to lexical properties, which gives us the actual measurement of the emotion’s intensity. The sum of intensities of each word in the text data gives the sentiment score for the text input.

Let’s understand this with an example. Optimism is conveyed by using all of these words “enjoy”, “like”, “love”, and “happy”. And, the actual meaning of these words or phrases which include “did not love” is understood by the VADER. VADER also has the capability to recognize how important casing of letters and punctuations are.

So, VADER is used for text analysis.

First the title field is examined and the compound score is calculated by this VADER sentiment intensity analyzer from the NLTK’s library. This score resembles the overall sentiment of the text. This compound score is just a single number which has the boundary from 1 to -1. The compound score of 1 indicates a rare optimistic outlook and the compound score of -1 shows an extremely negative emotion.

```
In [25]: body_data.head()
```

Out[25]:

	body	timestamp	Positive Sentiment	Neutral Sentiment	Negative Sentiment	# Of Words	# Of StopWords	Average Word Length	# Of Times Events Was Mentioned	# Of Organizations Mentioned	# Of Political locations Mentioned	# Of Non Political locations Mentioned
0	ifrom the us historically not politically acti...	2024-04-10	0.000001	0.685001	0.315001	23	10	5.692308	0	0	1	0
2	like trade on land water and air and air travel	2024-04-10	0.217001	0.783001	0.000001	10	4	4.333333	0	0	0	0
3	why is the uae so hostile towards algeria rece...	2024-04-10	0.000001	0.877001	0.123001	25	10	7.133333	0	0	3	0
5	he is currently hiding in qatar and his commen...	2024-04-10	0.000001	0.821001	0.179001	25	12	7.000000	0	0	1	0
9	x200b	2024-04-10	0.000001	1.000001	0.000001	1	0	5.000000	0	0	0	0

In order to compute the compound score, the valence scores of each word from the text data are added and standardized between the range between +1 and -1. Their individual polarities and the environment in which the words are utilized is considered.

Compound rating of ‘0’ indicates a neutral sentiment. Nevertheless, values greater than ‘0’ and lesser than ‘0’ indicate positive and negative sentiment, respectively. Overall intensity of a specific sentiment is denoted by the compound score's magnitude, in which greater scores indicate greater feeling. Therefore, summary of the overall sentiment of the text data is given by both context and polarity of each word which are considered for the compound score by the sentiment intensity analyzer.

Sentiments are classified according to the compound sentiment scores :

- **Positive Sentiment :** Compound score > 0.05 (This value may change for different contexts).
- **Neutral Sentiment :** -0.05 <= Compound score <= 0.05.
- **Negative Sentiment :** Compound score < -0.05 (This value may change for different contexts).

NAIVE FEATURE EXTRACTION

The following feature extraction techniques have been applied to the 'body' column in the 'body_data' DataFrame :

- **Number of Words (# Of Words) :** The 'body' column has been processed to calculate the total number of words in each text entry. This feature provides insight into the length or complexity of the text.

- **Number of StopWords (# Of StopWords) :** Stop words, common words like "and," "the," etc., which may not contribute significant meaning, have been identified and counted for each text entry in the 'body' column. This feature is useful for understanding the prevalence of common language constructs.

- **Average Word Length (Average Word Length) :** The average length of non-stop words in each text entry has been computed. This feature offers an indication of the typical length of words used, potentially reflecting the complexity or style of the language.

NAMED ENTITY EXTRACTION FEATURES

- **Frequency of events being referred to (the number of times events are mentioned) :** For each text item in the 'body' column, the code counts the instances of entities tagged as 'EVENT' using the named entity recognition model. This function provides information on how frequently events are mentioned in the text.

- **Amount of organizations being referred to (# Of Mentioned) :** For each text input, the code counts the instances of entities labeled as 'ORG'. This feature indicates how often an organization is referenced in the text.

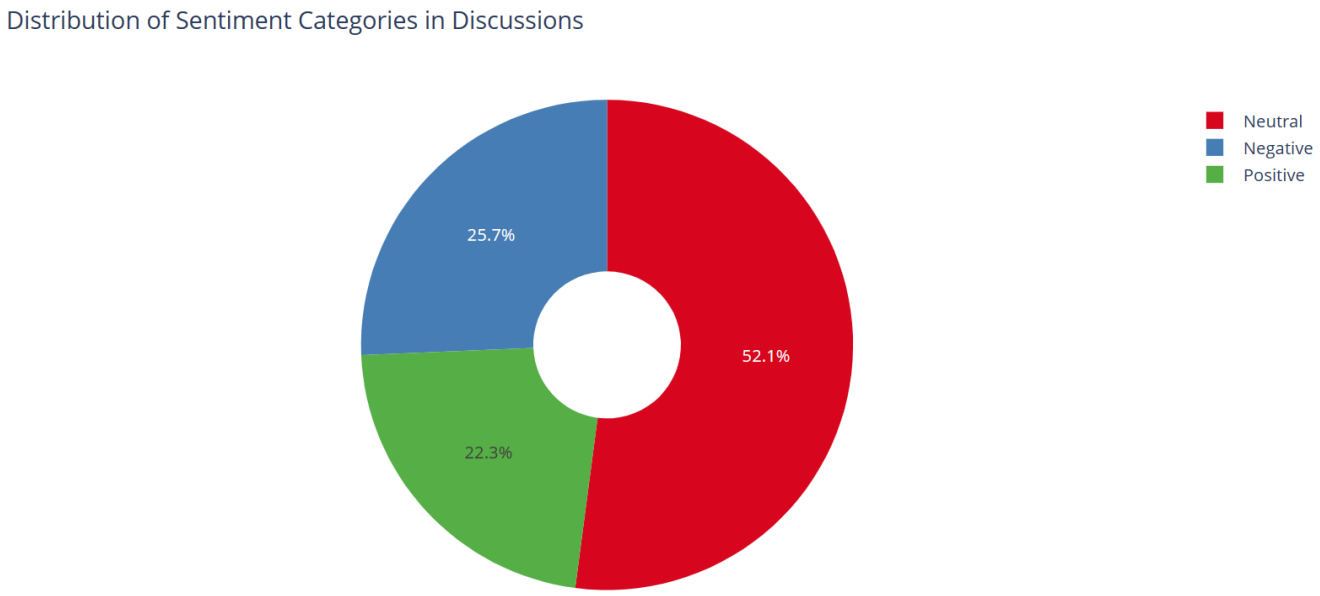
- **Amount Political Locations referred to (# Of Mentioned Political Locations) :** 'GPE' (geopolitical entities) are identified and counted by the code for every text entry. This feature highlights how frequently political topics are referenced.

- **Listing of Non-Political Places referred to (# Of Non-Political Places Mentioned) :** 'LOC' Entities (areas) are counted for each text input, providing data on how frequently nonpolitical locales are referenced.

The Reddit talks dataset is processed and features are engineered through a series of processes that concentrate on sentiment, linguistic, temporal, and statistical characteristics. The result features help with text analysis, sentiment pattern detection, and data preparation for further analysis or model development.

RESULTS AND ANALYSIS

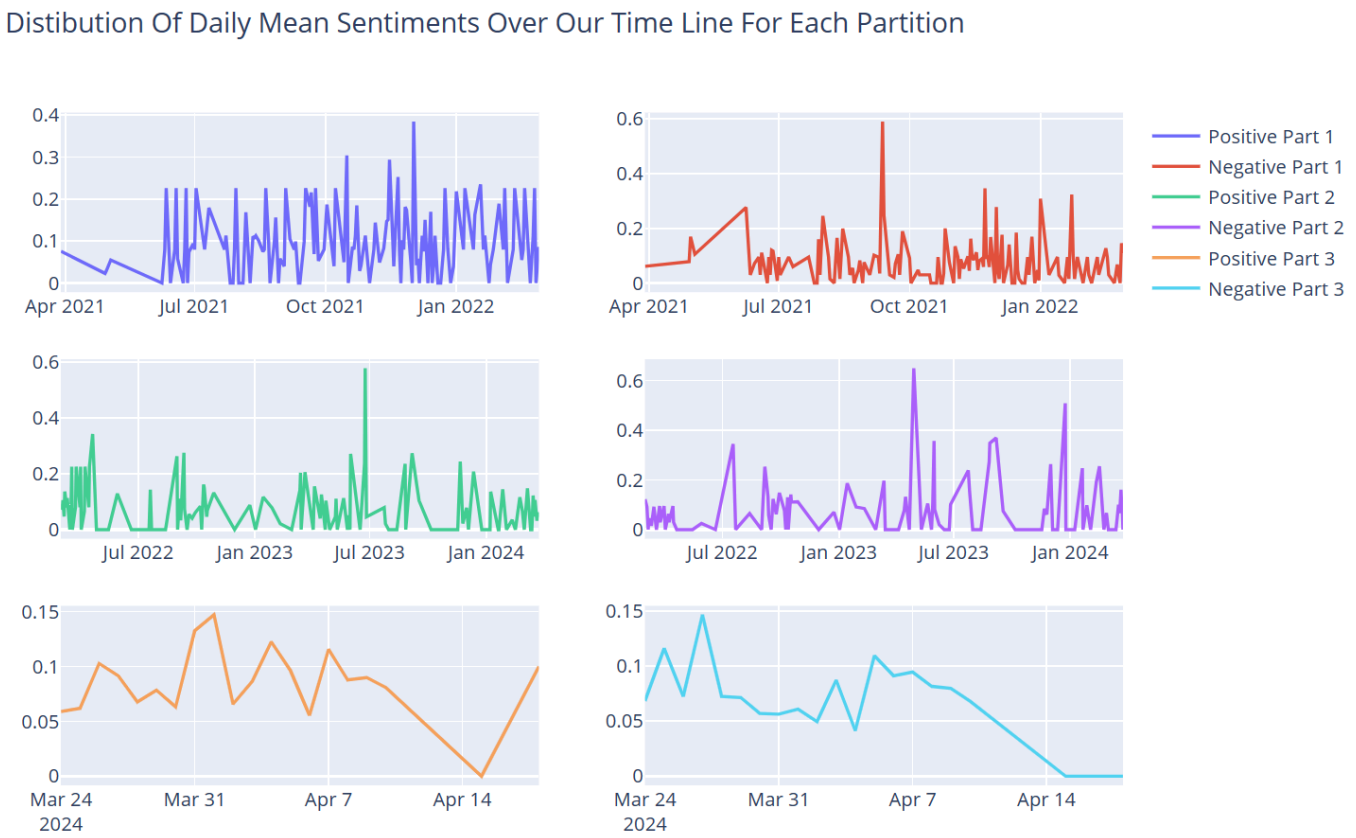
SENTIMENT DISTRIBUTION



The above distribution represents the percentage of discussions with an unfavorable sentiment score (compound score < -0.5), favorable sentiment score (compound score > 0.5), and neutral sentiment score (-0.5 < compound score < 0.5). Based on the information shown above, we conclude that 48.4% of the conversations are negative, or nearly half of the total. This also suggests that citizens in both nations want peace, opposing violence. Additionally, there were no doubts that some people who benefited from these disputes and some derogatory language were disguised as favorable.

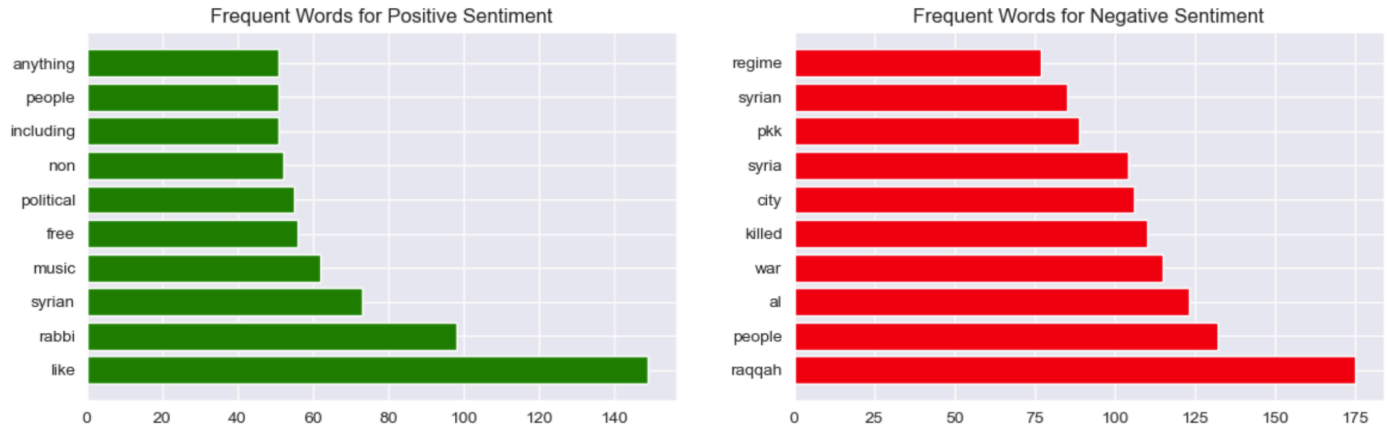
SENTIMENTS OVER TIME

Sentiment analysis relies heavily on long-term analysis to identify trends and anticipate future events. Three partitions may be seen in the plot above: feelings spanning years, months, and weeks. When we look at the weekly partition, we can see that, in terms of both positive and negative sentiments, there aren't many fluctuations in the mean average of the respective sentiments up until October 8th. However, once the Syria war broke out on October 7th, 2023 (Reference), we can see that there are fluctuations, indicating that people have been talking about it more since then. These fluctuations would be the result of people's opinions changing and more conversations starting after October 7th. This sentiment analysis finding is quite helpful. The benefit may be that, in the future, we will be able to identify impending events and take appropriate action if we observe a sudden shift (fluctuation) in these types of opinions in real time analysis.



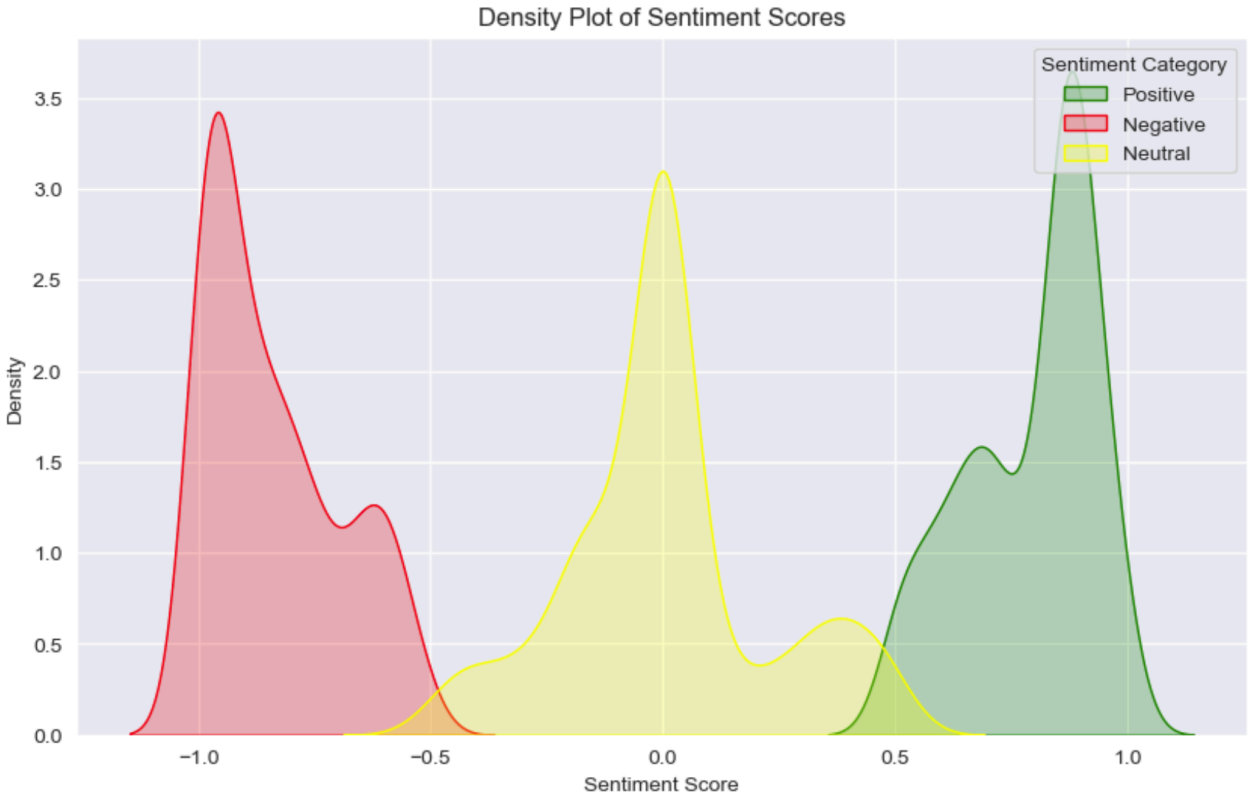
FREQUENCY OF WORDS FOR SENTIMENTS

The positive and negative analysis got the results and we have taken most terms in both the analysis. These were obtained by tokenizing the words, deleting stop words, and then calculating word frequency.

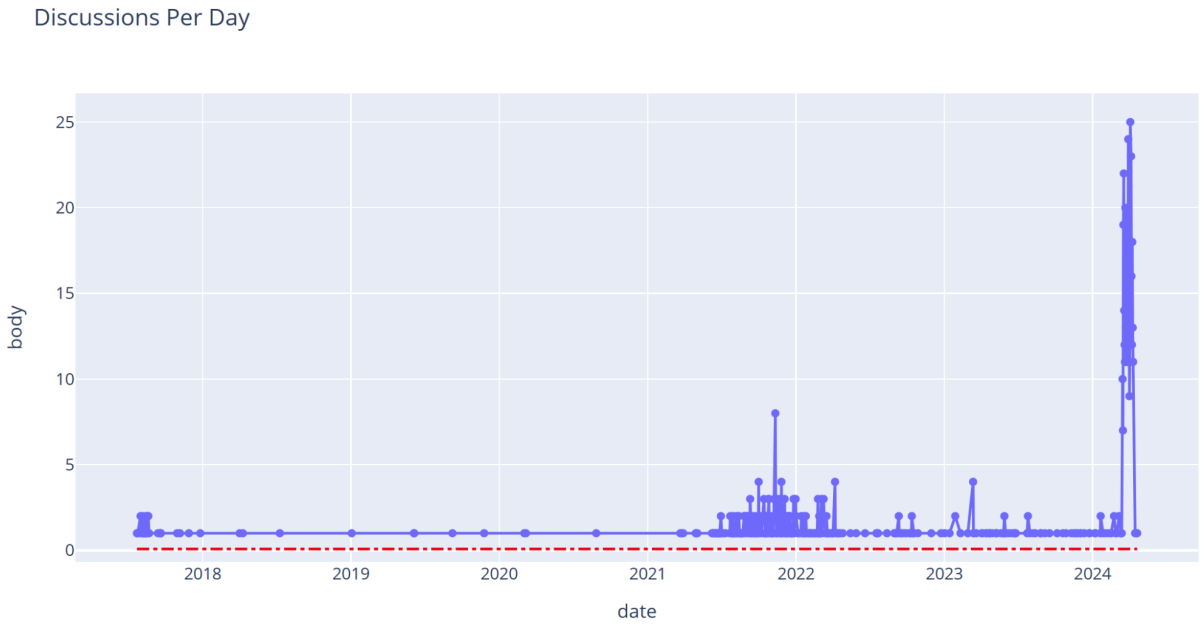


DENSITY PLOT OF SENTIMENT SCORES

The density graphs shown above allow us to conclude that there was a significant quantity of unfavorable sentiment. This result is in line with the distribution map, which indicates that there was high density in 48.4% of the talks with unfavorable complex sentiment scores (negative category). But looking at the negative category's peak shows that most feelings were unfavorable with the density curve explaining the remaining sentiments.



FREQUENCY OF DISCUSSIONS OVER TIME

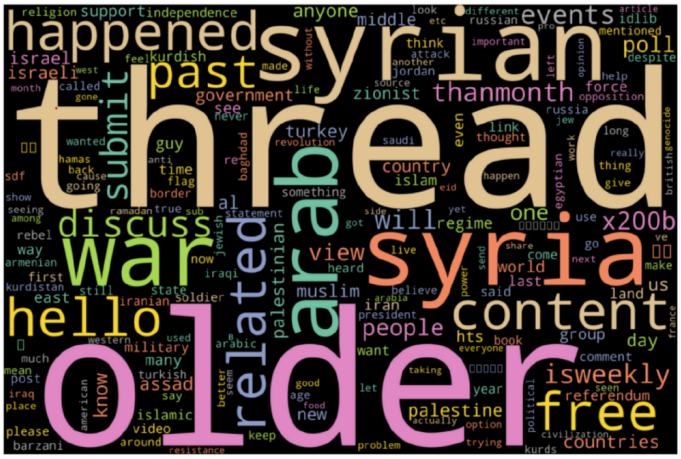


Peaks in conversation frequency can be seen near MAY 2021 (Reference[1]) and OCTOBER 2023 (Reference[2]), when we plot the frequency of discussions over time. Based on the two connections provided, it may be inferred that the peak dates for severe conflict tensions between the two countries are May 10, 2021, and October 8, 2023. These dynamic charts above allow for considerably more code-based validation of the results.

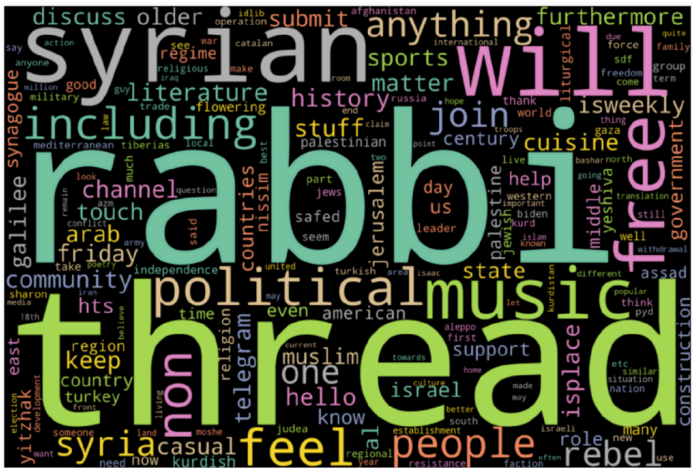
WORD CLOUDS

Compound Data

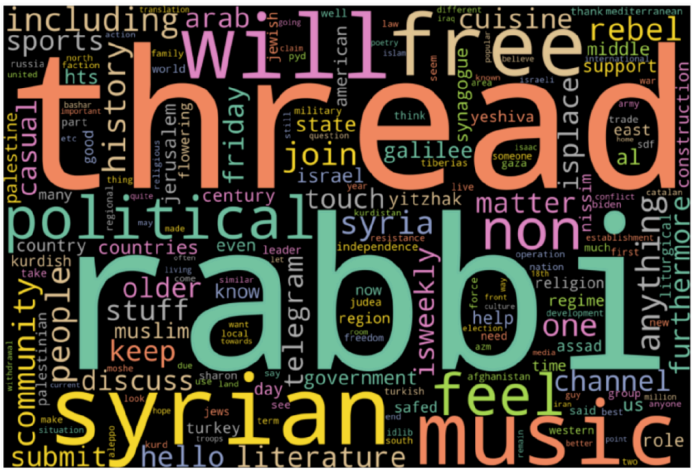
Neutral Data



Positive Data



Negative Data



- **Whole Data :** Word cloud has all the data which we have collected, with no sentiment filter applied. The most common terms in the dataset, such as "Syria," "Hamas," and "peace," indicate that they are important subjects or entities. The use of such terms suggests that political and conflict-related subjects would probably be the main focus.

- **Neutral Sentiment Data** : Phrases in this cloud are presumably gathered from text data deriving as having an unbiased sentiment toward the conflict. Certain terms used here, such as "people," "Syria," and may allude to debates or comments that do not endorse or condemn any position, but rather bring them up in an unbiased manner.

- **Positive Sentiment :** Material with a positive attitude toward the problem is shown in this word cloud. Words that imply hope or support, such as "peace," "state," and “will," may imply positive or support for the trend of comments which are taking part in the media.

- **Negative Sentiment :** Text containing a negative sentiment is represented by this word cloud. Phrases like "war," "conflict," and "attack" may be used to express disapproval, criticism, or worries about the things which are going on with trends such as threads.

With the use of these word clouds, one may rapidly identify the primary themes within lengthy text passages and ascertain the general attitude of a dataset toward a given subject. They are frequently employed in

sentiment analysis, a widely used data science and natural language processing activity that aims to understand trends in textual data, social media debate, and public opinion.

CONCLUSION

- **Effective Use of PRAW :** The project made good use of the PRAW library to gain access to Reddit's API, which enabled the thorough gathering of a variety of data from communities that were involved in conversations around the Syrian conflict.

- **Careful Data Preprocessing :** By addressing issues like duplication and missing values, issues with data preprocessing were resolved and the acquired data was made ready for further analysis.

- **Using NLTK's VADER Tool :** To gain a thorough grasp of public sentiments, the project employed NLTK's VADER sentiment analysis tool in order to get nuanced sentiment scores— positive, negative, neutral, and compound—across the talks.

- **Feature Engineering for Depth :** Feature engineering was essential in giving the sentiment analysis more depth by combining a number of variables such as textual characteristics, temporal features, statistical metrics, and word frequency analysis.

IMPLICATIONS AND FUTURE DIRECTIONS

- **Significance of Results :** By considering the Syrian conflict with the major trends, the project's results offer us important considerations for academics, policymakers, and everyone else.

- **Acknowledgment of Difficulties :** Well the best way is to know what we have faced at the time, the issues about data quality and any possible prejudices in sentiment analysis, to ensure a fair and transparent evaluation of the project's constraints.

- **Future Research Directions :** By creating the best-known paths for additional research which include specific sentiment patterns, the study contributes to the continuing discussion on sentiment analysis in geopolitical contexts.

- **Thoughts on Technique :** The technique that we selected was very positive, for the wellbeing of the tools we have considered the effective mindset was used, and suggestions for improvements in future projects adds a degree of comprehension to the task's overall process.

- **Expanded Role of Sentiment Analysis :** The study goes beyond its main motive to highlight the crucial role of sentiment analysis when analyzing public opinions on how they react to major topics. It also highlights how important sentiment analysis is to understand what people have to say on social media.

REFERENCES

- https://en.wikipedia.org/wiki/Syrian_civil_war
- https://www.britannica.com/event/Syrian-Civil-War