Sentiment Analysis of Palestinian-Israeli Conflict

Yijie Bai, Zhisheng Jin, Tabib Wasit Rahman

1 Introduction

1.1 Background

The Palestinian-Israeli conflict has been a constant state of affairs in the international arena. Recent years have seen significant developments in the Israeli-Palestinian conflict, with increasing concerns about escalating tensions. [1]

Escalation and Humanitarian Challenges: The Israeli-Palestinian conflict is increasingly acute, reaching a critical point amid ongoing violence, illegal settlement expansions, and a stalled peace process. This escalation is marked by attacks on civilians, increased use of arms, and settler-related violence, including recent bombings in Jerusalem and violent incidents in Hebron. In Gaza, the fragile situation has been further exacerbated by militant activities and Israeli airstrikes. Despite efforts by the UN to mediate ceasefires and improve local conditions, such as providing fuel for Gaza's power plant and assisting needy families, restrictions continue to impede humanitarian and development work.

Pursuing Peace and Political Solutions: The UN envoy underscores the urgent need for a two-state solution, still supported by many Palestinians and Israelis. To move forward, it's essential to engage with both parties to reduce tensions, counter negative trends affecting final status issues, and improve access and trade to strengthen the Palestinian economy. Additionally, reinforcing Palestinian institutions and governance is crucial, including implementing democratic reforms, conducting elections in the Occupied Palestinian Territory, and ensuring the effectiveness of Palestinian security forces. These steps necessitate a formidable international commitment and coordinated, sustained attention to the conflict.

The main points of contention in the Palestinian-Israeli conflict include:
1) Territorial disputes: The two sides disagree over the ownership of areas such as Jerusalem and the West Bank.
2) Religious beliefs: Judaism, Christianity, and Islam all regard Jerusalem as a holy site, and as a result, contests over the area often lead to religious conflict.
3) Political strife: The Palestinian National Liberation Movement (PLO) and other Palestinian political organizations seek the establishment of an independent Palestinian state, while Israel favors the status quo or further territorial expansion.

In recent years, the international community has been working to advance the Palestinian-Israeli peace process, including the United States, the United Nations, and other countries. However, due to many complex factors, peace in this region remains elusive. Public views on the Palestinian-Israeli conflict also vary. [2]

The goal of this project is to utilize the Sentiment Analysis technique from Natural Language

Processing in order to visualize the public's opinion and sentiment regarding the Palestinian-Israeli conflict. As part of our effort to achieve this goal, we collect data from Reddit, specifically the "worldnews" subreddit, which contains a large number of posts regarding this subject. In order to identify positive or negative sentiment, we annotate the data according to whether the post supports or opposes the Palestinian-Israeli Conflict. We aim to answer the following research questions:

- RQ1: What sentiment do Reddit users generally express about discussions of Palestinian-Israeli Conflict?
- RQ2: What are the topics associated with the positive and negative sentiments?
- RQ3: What are the dominant features in the discussions of Palestinian-Israeli Conflict?

There is an abundance of textual data available on social media platforms, allowing researchers to gain access to public opinion in real time. As a result of these platforms, human behavior and opinions regarding major global issues have been studied. We use a web crawler to collect relevant posts, comments or tweets from a subreddit called “world news”.

Finally, the results are analyzed, and the results of the data analysis are transformed into understandable information. We use visualization tools such as charts, hot words, etc. to present the results of the analysis.

1.2 What is the NLP?

NLP began in the 1950s as the intersection of artificial intelligence and linguistics. NLP was originally distinct from text information retrieval (IR), which employs highly scalable statistics-based techniques to index and search large volumes of text efficiently.[3]

Natural Language Processing (NLP) is a critical field within artificial intelligence that focuses on the interaction between computers and human language. It aims to equip computers with the ability to understand, interpret, and respond to human language in a way that is both intelligent and meaningful. This involves teaching machines to comprehend the intricacies of human language, including its syntax (the structure of sentences), semantics (the meaning of words and phrases), and pragmatics (the context in which language is used). NLP is a complex domain, as human language is inherently nuanced and varied, featuring elements like idioms, slang, dialects, and cultural references.

Natural language processing is rapidly growing in popularity in a variety of domains, from closely related fields like semantics [4], linguistics (e.g. inflection, phonetics and onomastics, automatic text correction), and named entity recognition [5] to distant ones like bibliometry, cybersecurity, quantum mechanics, gender studies, chemistry, and orthodontia [6].

The applications of NLP are diverse and far-reaching, impacting a wide array of technologies and industries. Key applications include language translation, sentiment analysis, speech recognition, and the development of chatbots and virtual assistants like Siri and Alexa. Moreover, NLP is instrumental in text analytics, which involves extracting useful information from large volumes of text, such as emails, web documents, and social media posts. The field has

evolved significantly with advancements in machine learning and AI, particularly with the advent of deep learning models like Transformers, which have brought substantial improvements in language understanding and generation. As such, NLP stands as a cornerstone technology in our digital era, continually bridging the gap between human communication and computerized data processing.

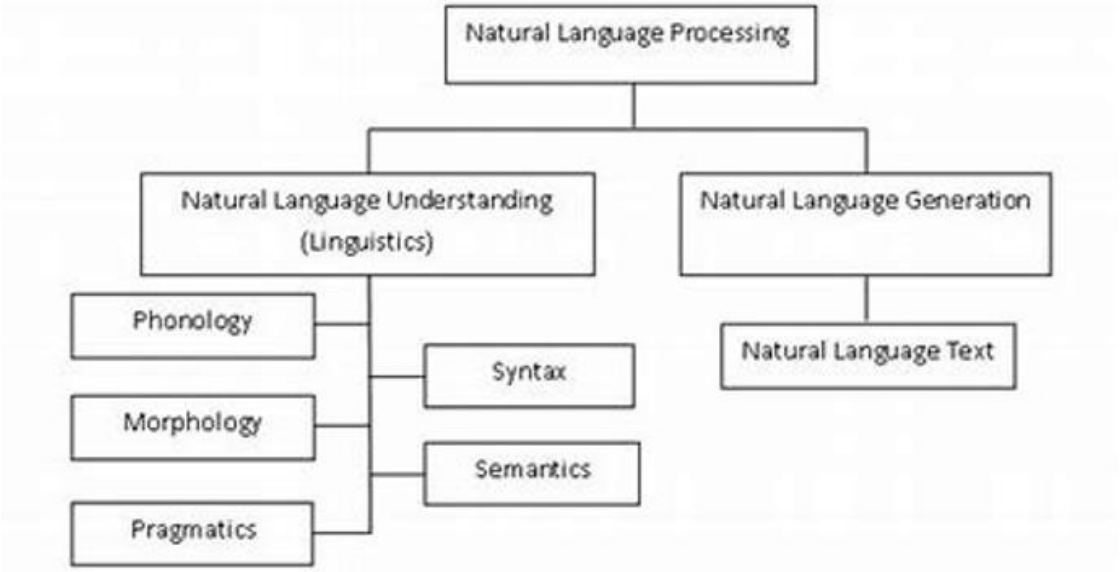

Fig. 1 Natural Language Processing steps

2 Related Work

2.1 Sentiment Analysis

Sentiment Analysis, also known as Opinion Mining, is a subset of Natural Language Processing (NLP), text analysis, and computational linguistics. It involves identifying and classifying subjective information in text to understand and interpret the emotional tone conveyed. The primary task in sentiment analysis is to categorize text into positive, negative, or neutral sentiments. In more advanced scenarios, it can also detect specific emotions like anger, happiness, or sadness. This analysis is widely applied across various domains, including market analysis, social media monitoring, product reviews, and public sentiment tracking.

The techniques used in sentiment analysis encompass lexicon-based approaches, machine learning methods, and deep learning strategies, each capable of processing diverse types of textual data such as social media posts, reviews, and news articles. Challenges in this field include recognizing sarcasm, idioms, and expressions that aren't meant to be taken literally, as well as dealing with cross-cultural and cross-linguistic variations in emotional expression. Sentiment analysis holds significant commercial value, as businesses and organizations leverage it to gauge consumer sentiment towards brands, products, or services, thereby informing marketing strategies, product development, and customer service. As AI and machine learning technologies continue to advance, the accuracy and applicability of sentiment analysis are steadily expanding.

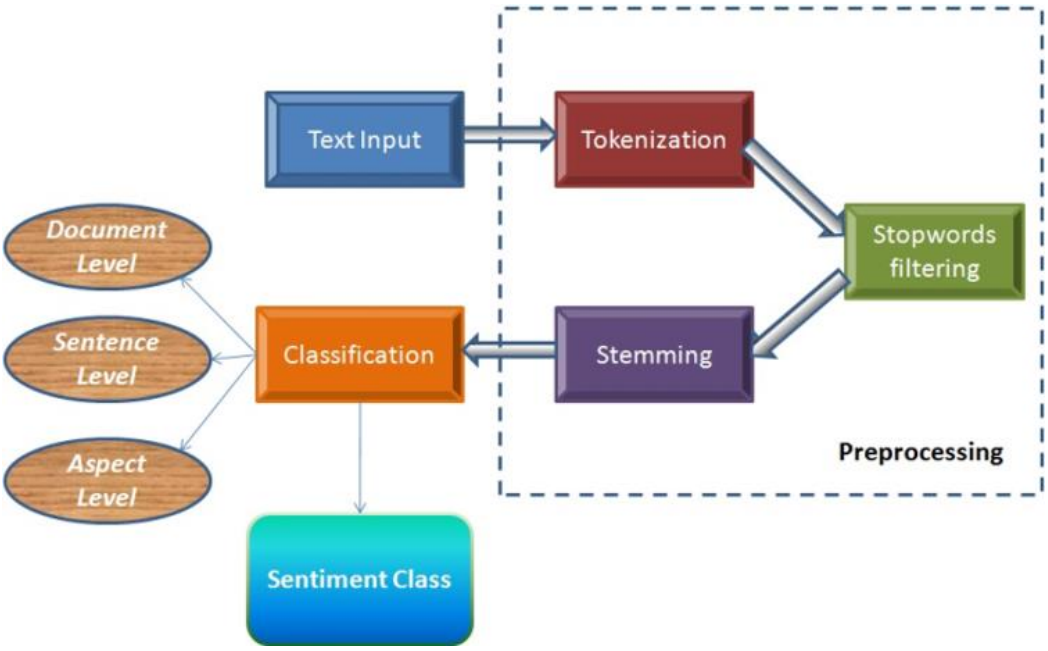

Fig. 2 Sentiment Analysis steps

2.2 Israel-Palestine conflict

A Pew Research Center study provides insight into American public opinion regarding the Israel-Hamas conflict as of late 2023. [7]

Responsibility Attribution: A considerable majority of Americans (65%) perceive Hamas as bearing substantial responsibility for the ongoing conflict, in contrast to 35% who attribute significant responsibility to the Israeli government. A smaller proportion of respondents assign considerable blame to the Palestinian (20%) and Israeli people (13%). This viewpoint demonstrates notable differences along party lines, with 73% of Republicans and 62% of Democrats considering Hamas highly responsible, while Democrats are more than twice as likely as Republicans to hold the Israeli government equally accountable.

Concerns and Government Response: The conflict has raised concerns among nearly half of Americans (48%) about potential increases in violence against Jews in the U.S. This apprehension is shared equally by Democrats and Republicans. Regarding the Biden administration's handling of the conflict, opinions are divided: about 35% approve, while 41% disapprove of the administration's response. This division is particularly pronounced among younger adults under 30, where 46% disapprove. Americans also have mixed views on whether President Biden is favoring either side or maintaining a balanced approach.

Military Operations and Two-State Solution: Public opinion is split on Israel's military operations against Hamas. Around 27% believe Israel is going too far, whereas 25% view it as an appropriate response, and 16% think Israel isn't going far enough. Democrats are more likely than Republicans to believe Israel is overreaching. About half of the Americans (52%) still see a two-state solution as a viable future outcome, with a higher proportion of Democrats than

Republicans holding this belief. Engagement with news about the conflict varies, with older adults following more closely than younger ones. Those more informed tend to have stronger opinions about the administration's response to the conflict.

A recent study on Reddit's role in the 2016 U.S. Presidential elections revealed insights into online political interactions, challenging the common narrative of echo chambers. Contrary to the expectation of like-minded individuals reinforcing each other's beliefs, the study found a preference for cross-cutting political interactions, particularly among Clinton supporters responding to Trump supporters. This indicates an asymmetrical interaction pattern between the opposing groups. Additionally, the study noted a trend of geographical homophily in online interactions and a small positive correlation between cross-interactions and voter abstention, suggesting that exposure to opposing views might relate to decreased political participation. These findings underscore the complexity of public opinion formation on social media, highlighting the coexistence of polarization with a significant presence of diverse interactions. This challenges the conventional approach of addressing echo chambers and emphasizes the need for a deeper understanding of the underlying social dynamics before implementing potential technical solutions. [8]

Analyzing public opinion encompasses a range of traditional and modern methods. Traditional approaches include surveys and questionnaires, focus groups, case studies, and field observations, which provide direct responses and in-depth insights but can be time-consuming and limited in scope. Modern methods leverage technology, with social media analysis offering real-time, large-scale sentiment and trend assessment. Text analysis and Natural Language Processing (NLP) extract themes and sentiments from unstructured data like news and blog posts. Data mining and machine learning reveal patterns and correlations in large datasets, while web behavior analytics uncover online preferences and trends. Sentiment analysis evaluates emotional tones in language, useful for scrutinizing social media and online comments. Lastly, big data analysis handles complex, voluminous data sets to glean insights about public opinion, suitable for multifaceted data sources. The combination of these diverse methods often yields comprehensive and nuanced analysis results.

3 Methodology

In the field of sentiment analysis, there are multiple methodologies that can be applied, each with their own characteristics and scenarios of applicability. Here are some common sentiment analysis methodologies and how they compare to each other:

a. Lexicon-based and Rule-based Methods:
Lexicon-based approaches to sentiment analysis rely on predefined emotion dictionaries, such as AFINN or SentiWordNet, making them easy to implement without the need for training data. However, they often fall short in accurately capturing contextual meanings and struggle with interpreting sarcasm or idiomatic expressions. Similarly, rule-based methods use specific rules to identify emotions, tailored for particular scenarios. While effective in certain contexts, they lack flexibility and may not adapt well to the diversity and complexity of language.

b. Machine Learning and Deep Learning Methods:

Machine learning methods, including algorithms like Naive Bayes and Support Vector Machines (SVM), are capable of handling complex linguistic features and are suitable for large-scale data. The downside is their dependency on extensive labeled training data and potential inadequacy in dealing with unseen data or new domains. On the other hand, deep learning approaches, using Recurrent Neural Networks (RNN) or Long Short-Term Memory networks (LSTM), excel in extracting features and capturing deep semantic structures. Despite their powerful feature extraction capabilities, they require substantial amounts of data and computational resources, and often suffer from a lack of interpretability.

We mainly analyze the titles and comments. We try to analyze these texts, each piece of data (title and text) of this data for this Palestine-Israel conflict, what kind of a view it is, whether it is positive or negative, and ultimately we can determine that the vast majority of Internet users on this site are positive or negative.

Initially, data collection is done using Reddit API or third-party scraping tools, focusing on specific subreddits or keywords to ensure data relevance. This is followed by data preprocessing, which includes cleaning the data to remove irrelevant or duplicate content and text preprocessing tasks such as removing stopwords, punctuation, numbers, and special characters, and performing stemming or lemmatization. Text tokenization might also be necessary, especially for non-English texts.

Feature extraction is the next step, where text is transformed into a format that machine learning models can process, commonly through methods like Bag of Words, TF-IDF, or word embeddings like Word2Vec.

The process culminates in data analysis and visualization, where sentiment analysis results are examined for trends, patterns, or significant differences, and tools like charts or word clouds are used for visualization to better understand and present the data. This methodology requires researchers to have the appropriate technical, analytical, and ethical knowledge and skills, with a focus on ensuring data quality and correctness of the analysis.

3.1 Data Collection

We crawled a subreddit called “world news”. While crawling the data, we stored every 5000 rows of data we crawled. We noticed that if we crawled too much data at once, it would error out and just stop us from crawling the data because of the frequent access. To avoid this, we rowed down many small datasets by constantly storing small datasets, which we eventually merge into one large dataset.

3.2 Data preprocessing

Data cleaning: removed irrelevant content such as advertisements, spam, etc.
Text preprocessing: including removal of deactivated words, stemming extraction, lexical labeling, etc.

Fig. 3 Data Cleaning

As a first step, we de-weight, and then perform natural language processing by removing invalid characters such as punctuation marks as well as some of the most commonly used words (words that do not convey subjective judgments or personal feelings), such as some pronouns. Neither the positive nor negative judgments associated with these data will affect the analysis of the project.

To clean up the data, the following methods were used: Delete non-alphabetic characters. Converted all characters to lowercase. Removed English stop words from the "nltk" module. Set missing text fields to "unknown". Set missing numeric fields to zero

Moreover, there are also posts like some that are titled as a comment on the original post. To maintain the hierarchy, as shown, the title of one of the comments (which was "Comment") is replaced with the title of the original post.

Table 1: Cleaned Dataset

Title	Text	Timestamp	Author	Upvotes	# of comments	URL
Belgian parliament refusing to screen video of...	Belgian parliament refusing to screen video of...	11/15/2023/15:24:30	TheRealJaneAusten	4	0	https://www.timesofisrael.com/belgian-parliame...
Iran's nuclear enrichment advances as it stone...	Iran's nuclear enrichment advances as it stone...	11/15/2023/15:15:42	GuiltySigurdsson	7	1	https://www.reuters.com/world/middle-east/iran...
Comment	So, uh, how do we know they're advancing if in...	11/15/2023/15:29:52	elFistoFucko	1	0	https://www.reuters.com/world/middle-east/iran...
Several bridges shut in Geneva due to flood th...	Several bridges shut in Geneva due to flood th...	11/15/2023/15:15:42	BezugssystemCH1903	6	1	https://www.swissinfo.ch/eng/society/several-b...
Comment	After heavy rain swept over Switzerland in t...	11/15/2023/15:16:26	BezugssystemCH1903	1	0	https://www.swissinfo.ch/eng/society/several-b...

Cleaned_Title	Cleaned_Text	Log_Upvotes	Log_#_of_Comments	Sentiment	Sentiment_Label	Topic
belgian parliament refusing screen video hamas...	belgian parliament refusing screen video hamas...	1.609438	0	0	Neutral	11
iran nuclear enrichment advances stonewalls un...	iran nuclear enrichment advances stonewalls un...	2.079442	0.693147	0	Neutral	11
iran nuclear enrichment advances stonewalls un...	uh know advancing inspectors allowed inspect b...	0.693147	0	-0.025	Negative	11
several bridges shut geneva due flood threat	several bridges shut geneva due flood threat	1.94591	0.693147	-0.0625	Negative	10
several bridges shut geneva due flood threat	after heavy rain swept switzerland past days...	0.693147	0	0.12244	Positive	10

3.3 Unsupervised topic extraction

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which

each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.[9]

From the dataset, we have a list of titles; we train an LDA model on the titles to identify topics and their associated terms.

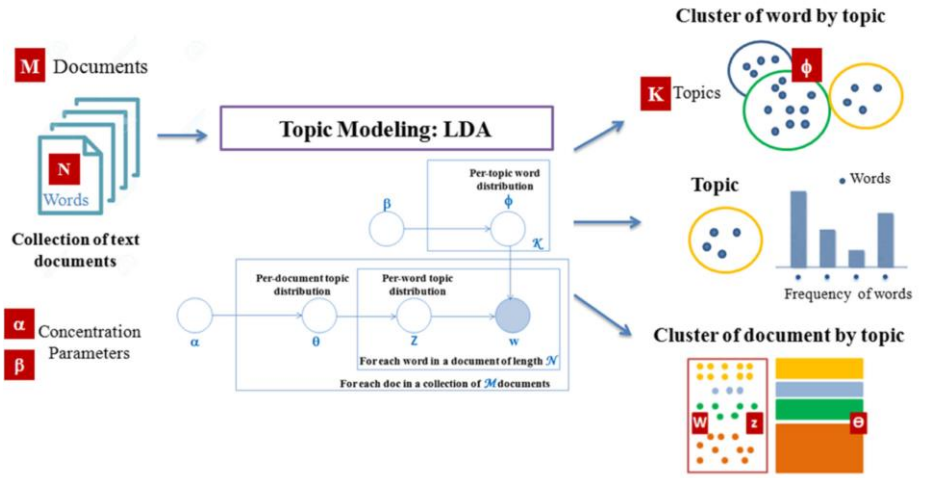

Fig. 4 LDA Model

3.4 Feature Extraction

We vectorized the title and text of each post using CountVectorizer and obtained the best fit parameters by fitting the text of each post with the number of upvotes using GridSearchCV. Then we created a RandomForestRegressor using the best-fit parameters and extract important features from the best-fit RandomForestRegressor

3.5 Classification using XLNet

To gain an insight on the general sentiment towards both sides of the conflict, a transformer model was fine-tuned to classify posts we collected in terms of he benefactive of the event the title is describing. Essentially, a general pattern of a post in r/worldnews are news title styled texts that describes an event being reported. Thus, aggregated sentiment in the comment section in one post reflects the general sentiments the users have for the event. By evaluating the benefactive of one event and the general sentiment of it, we can utilize the structure of 'post-comment' which is extremely common in social media and exempt us from redundant computation induced by treating post and comments as independent data points.

XLNet(Yang et al. 2019) is one of the extensions of the Transformer-XL model which takes the advantage of an autoregressive method to capture the bi-directional nuance of the context. It outperformed BERT in various benchmarks such as text classification and sentiment analysis, which is great for our objective.

We manually annotated 106 titles randomly sampled from our corpus in terms of whether it describes an event that is in favor of Israel, Palestine, or just Neutral. By splitting the dataset into training set, validation set, and test set, we trained a base-size pre-trained model on a linear

scheduled learning rate starting from 3e-5 in 10 epochs with 0.01 weight decay and an AdamW optimizer.

On the other hand, the algorithm we used to calculate the general sentiment in one post is that we calculate the sentiment of one comment using the aforementioned SVM we trained on the entire corpus, mapping it in the fashion of: *Positive : 1, Neutral : 0 and Negative : -1*. We iterate over the corpus once, and for each post we accumulate the sentiments of the comments in that post and weight them with their upvote counts. In one iteration we obtained the general sentiment of each post as an integer.

```python
# Sentiment mapping
def sentiment_to_numeric(label):
    mapping = {'Positive': 1, 'Neutral': 0, 'Negative': -1}
    return mapping.get(label, 0)

# One iteration over the whole corpus to get the general sentiment of each post
current_post_sentiment = 0
start_of_post_index = None
for index, row in df.iterrows():
    if pd.notna(row['Event_Benefactive']):
        if start_of_post_index is not None:
            df.at[start_of_post_index, 'General_Sentiment'] = current_post_sentiment
        start_of_post_index = index
        current_post_sentiment = 0

    current_post_sentiment += sentiment_to_numeric(row['Sentiment_Label']) * row['Upvotes']

# Handle the last post
if start_of_post_index is not None:
    df.at[start_of_post_index, 'General_Sentiment'] = current_post_sentiment

# Fill NaN values in 'General_Sentiment' with 8
df['General_Sentiment'].fillna(8, inplace=True)
```

Algorithm 1: Generate general sentiment by iterating through the corpus

4 Experiment

4.1 Datasets

Reddit, a popular social news aggregation and discussion platform, serves as a valuable data source for sentiment analysis due to several factors. First, its vast array of user-generated content, including comments and articles, offers diverse perspectives and emotional expressions, making it ideal for studying public opinion. Second, the platform's user base is diverse, encompassing a wide range of backgrounds and viewpoints, which enriches the discussion around specific events like the Israel-Palestine conflict. Third, Reddit's accessible data, facilitated by its API and third-party tools, allows for efficient data scraping and collection of substantial text data for analysis. Lastly, the textual nature of Reddit's content is well-suited for Natural Language Processing (NLP) techniques, such as sentiment analysis, to identify and extract emotional tendencies within the text. These aspects collectively make Reddit a rich resource for analyzing sentiments and opinions on various topics and events.

However, working with Reddit data may also present complications. Because of the myriad of media forms on Reddit, researchers may find that they need multiple methodological approaches in their analysis.[10]

Reddit posts, comments, and metadata can be accessed via the site itself, or via its APIs. Reddit's official API is free and publicly available and provides an array of functions. For these reasons, Reddit has an ecosystem of bots created by its user base to help in several ways, such as content moderation[11], adding functionality through summarizing information and linking to other websites, or providing humor through parody bot accounts.

The format of the data frame is illustrated by the following table.

Table 2: The dataset contains title, author, timestamp, and the number of upvotes and comments

Title	Text	Timestamp	Author	Upvotes	# of comments
Belgian parliament refusing to screen video of...	Belgian parliament refusing to screen video of...	11/15/2023/15:24:30	TheRealJaneAusten	4	0
Iran's nuclear enrichment advances as it stone...	Iran's nuclear enrichment advances as it stone...	11/15/2023/15:15:42	GuiltySigurdsson	7	1
Comment	So, uh, how do we know they're advancing if in...	11/15/2023/15:29:52	elFistoFucko	1	0
Several bridges shut in Geneva due to flood th...	Several bridges shut in Geneva due to flood th...	11/15/2023/15:15:42	BezugssystemCH1903	6	1
Comment	After heavy rain swept over Switzerland in t...	11/15/2023/15:16:26	BezugssystemCH1903	1	0

4.1.1 Some data related statistics

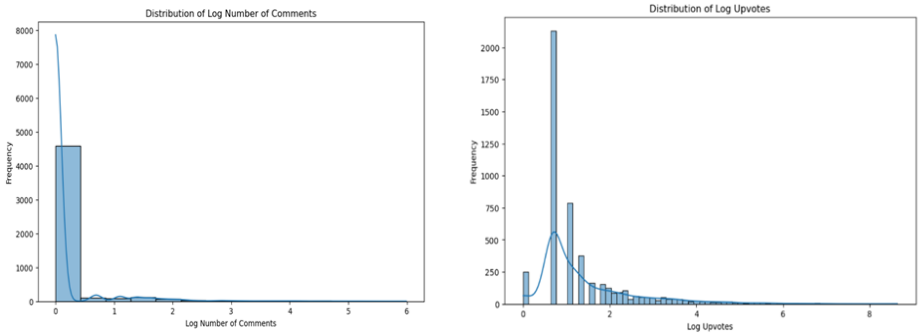

Fig. 5 Distributions of log number of comments and log upvotes

Table 3: descriptive statistics of upvotes and number of comments

	Upvotes	Number of Comments
count	5051	5051
mean	14.922986	0.919818
min	-144	0
25%	1	0
50%	2	0
75%	4	0
max	5613	399
std	135.117123	9.612765

4.2 Topic extraction and filtering

By fitting an LDA model on the whole corpus several topics are obtained. We restrained the number of topics to be at 15 due to empirical reasons. Topics unrelated to Palestine-Israel conflict were filtered out by a keyword search. Namely, we used the keywords shown in the following figure ("Palestine", "Palestinian", "gaza", "Israel", "Israeli", "zionist", "IDF", "Hamas") to filter out all the topics that do not contain any of them in their top 10 words.

```
def get_related_topic_indices(model, feature_names, no_top_words):
    related_topics=[]
    keywords= ['palestine', 'palestinian', 'gaza', 'israel', 'israeli', 'zionist', 'idf', 'hamas']
    for topic_idx, topic in enumerate(model.components_):
        top_words = " ".join([feature_names[i] for i in topic.argsort()[:-no_top_words - 1:-1]])
        print(f"Topic {topic_idx}:")
        print(top_words)
        if any(keyword in top_words for keyword in keywords):
            related_topics.append(int(topic_idx))
    return related_topics
```

Fig. 6 Function used for topic filtering

By using the pyLDAvis module we visualized the topic distribution of the whole corpus. We can see that in the Intertopic Distance Map there is a cluster of topics in the middle which are highly related to the Palestine-Israel conflict, which matches our expectation.

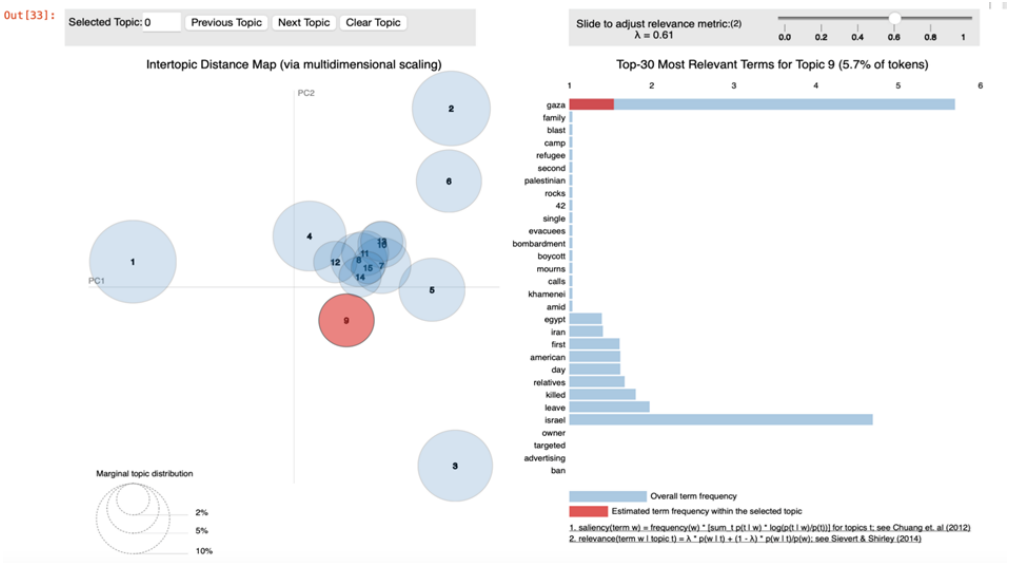

Fig. 7 LDA results visualization (λ = 0.61)

4.3 Feature Extraction:

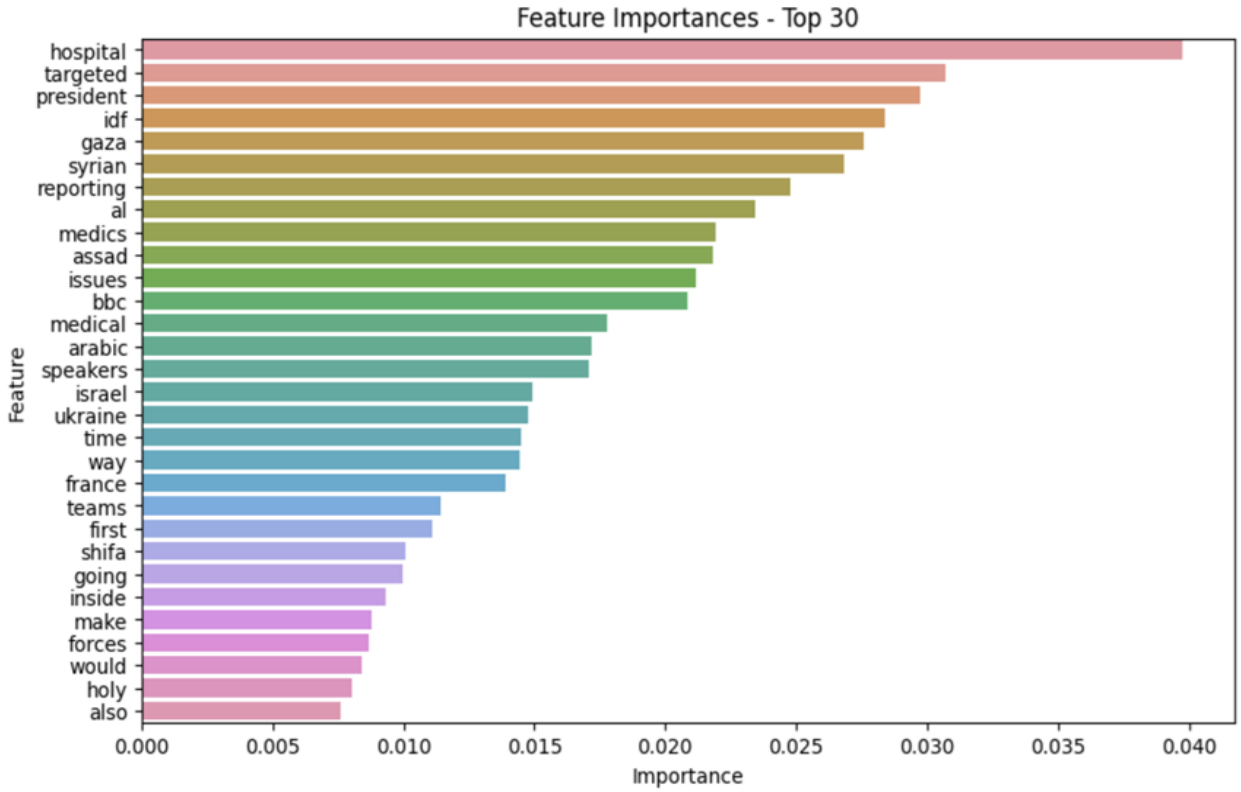

Fig. 8 Top 30 important features

4.4 Results:

a. Preliminary Results

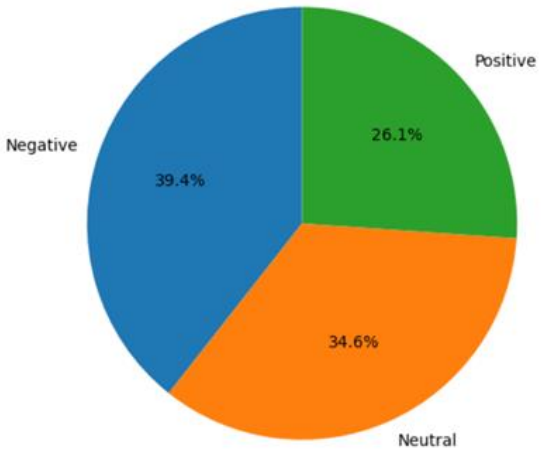

Fig. 9 Distribution of three sentiments across whole corpus

b. Customized SVM for measuring accuracy

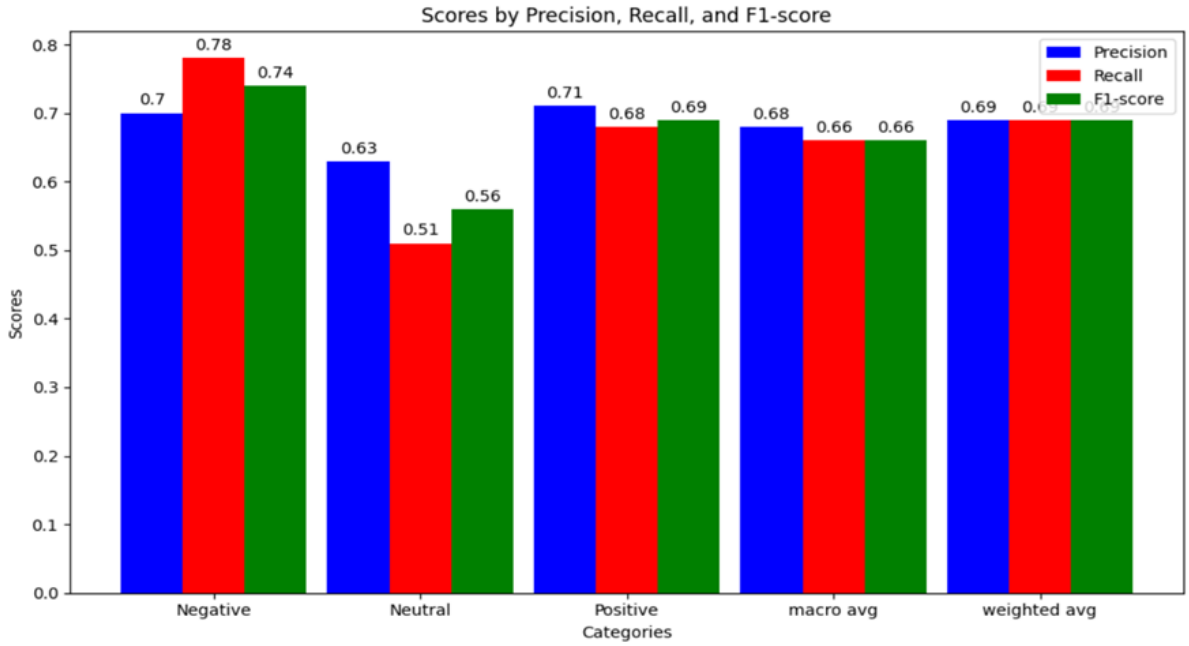

Fig. 10 Customized SVM sentiment classification performance

c. Top 10 Referenced Domain in Posts

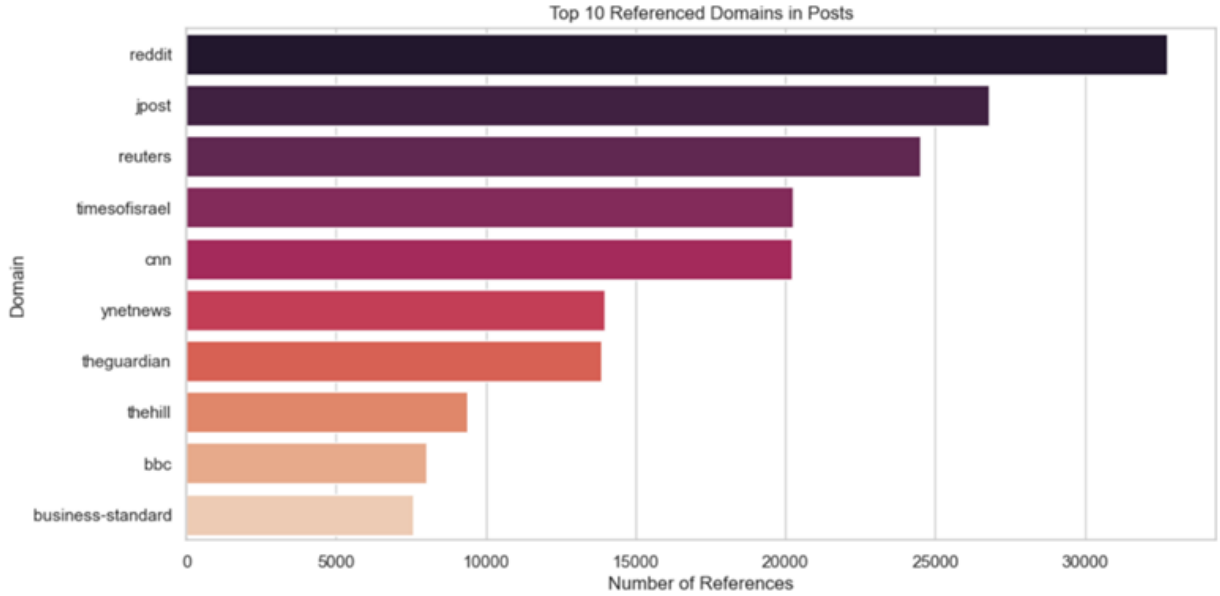

Fig. 11 Top 10 Referenced Domain in Posts

d. Topic cloud of dataset

Fig. 12 Top topics in a word cloud

e. XLNet classification training

The training results are shown in the following two figures:

Fig. 13 Training and Validation Losses Over Epoches

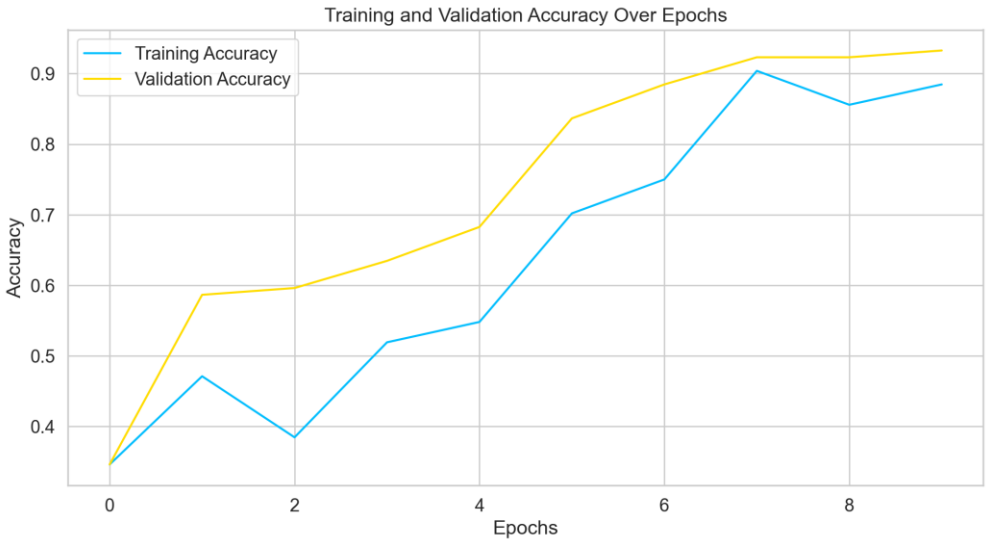

Fig. 14 Training and Validation Accuracy Over Epochs

By testing on the test set the model yielded an accuracy of 93.27%, detailed evaluation is shown in the following table:

Table 4: evaluation of the model's performance on the test set

category	precision	recall	f1-score	support
Israel	0.93	1	0.96	39
Neutral	0.91	0.97	0.94	33
Palestine	1	0.84	0.92	32
accuracy			0.94	104
macro avg	0.95	0.94	0.94	104
weighted avg	0.95	0.94	0.94	104

f. General sentiment results

The general sentiments for posts describing events in favor of either Palestine and Israel are shown in the following box plot. We capped the sentiment to 500 due to the presence of some outliers.

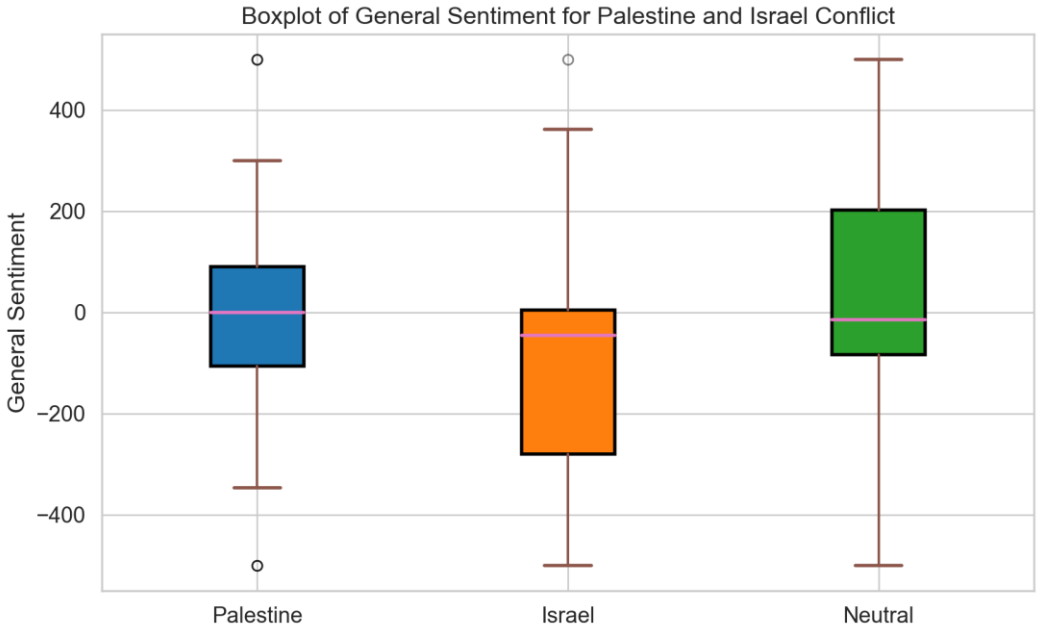

Fig. 15 Boxplot of the general sentiment for Palestine and Israel

From the plot we can see that the general sentiment towards events that are in favor of Palestine is nearly neutral. Israel on the other hand is skewed to the negative side. It's worth noting that most of the outliers are centered around Israel which can reach more than -30000. Since the general sentiment is just the number of negative comments weighted on the upvote numbers, these results show that in an environment in which neutral posts tend to yield positive comments, posts with an event in favor of Israel yields significant negative comments overall.

5 Conclusion

By analyzing the data set, people are mostly negative about the Palestinian-Israeli conflict, which is the first conclusion. The second conclusion is that we can see the distribution of data points through a box plot judgment, this distribution is very extreme, we found that there are a lot of people especially in favor of Israel, there are a lot of people especially in favor of Palestine (rather than a neutral attitude), at the same time, there are also some Palestinians who have a very positive attitude towards the war and some people are especially negative. Overall, the two conclusions of this project are that the vast majority of people are negative and pessimistic, and that there are many people who are more extreme in this area of public opinion.

When scraping data from Reddit for analysis, there are several challenges and considerations to bear in mind. Firstly, the data quality and potential biases present a concern. The Reddit community may not accurately represent the broader public opinion, leading to skewed data. Additionally, the nature of user comments, often laced with sarcasm, puns, or metaphors, can mislead Natural Language Processing analyses. [12] Secondly, privacy and ethical issues are paramount. Even though the comments are public, they might contain sensitive or personal information, necessitating adherence to privacy protection and data ethics standards. It's crucial to respect individual privacy and ethical considerations when analyzing personal opinions.

Technical and resource limitations also pose significant challenges. Scraping and processing large volumes of data require substantial technical resources, including storage and processing capabilities. Furthermore, Reddit's API has usage limitations, such as rate limits, which can impact the efficiency of data collection. As mentioned before, crawling too much data at once would be prohibited, so we took the approach of crawling in batches for final integration. Legal and platform policy compliance is another critical aspect. Users must follow Reddit's terms of use and policies, especially regarding data scraping and sharing, and be mindful of varying legal regulations across different countries and regions. Lastly, the complexity of data analysis should not be underestimated. The informal and dynamic nature of social media language presents challenges for NLP and sentiment analysis, necessitating advanced analytical skills and sophisticated algorithms to ensure the accuracy and reliability of the results. Therefore, while Reddit offers a rich data source, researchers need to carefully navigate multiple aspects, including data quality, privacy, legal, technical, and ethical issues, in their analyses.

Apart from Reddit, a multitude of other sources are available for data scraping, especially for social media and online content analysis. Twitter is a popular choice for tracking real-time events and public sentiments, with its API allowing access to tweets, user profiles, and metadata. Facebook, despite stricter data access regulations, remains a valuable source for analyzing public posts on pages and groups. Instagram, with its focus on images and videos, is ideal for scraping data related to visual content, including posts, hashtags, and user comments. LinkedIn offers insights into professional topics and industry trends, particularly for scraping career-related content and discussions. YouTube serves as a rich source for video content, comments, and user interactions.

Additionally, generic web scraping tools can gather text content from various websites and blogs, useful for acquiring information on specific topics or fields. News websites are crucial for

up-to-date event coverage and opinions. Online forums and discussion boards, like Stack Overflow and Quora, provide in-depth discussions on specific topics or questions. Review and rating sites such as Yelp and TripAdvisor offer user evaluations on products, services, and experiences. Moreover, many governments and public institutions provide open data sets, which can be utilized for diverse research and analyses. It's imperative to adhere to legal regulations and platform policies, particularly regarding privacy and data usage, when utilizing these sources.

Improving sentiment analysis of the Israel-Palestine conflict using Natural Language Processing (NLP) involves several key aspects. First, multilingual processing is crucial, as the region involves languages like Arabic and Hebrew. Understanding the context, including cultural, historical, and political nuances, enhances the accuracy of sentiment analysis. Incorporating multimodal analysis by combining text with video, audio, and images from social media offers a more comprehensive perspective.[13] Addressing and correcting biases in the analysis tools ensures a balanced view. Implementing a finer-grained classification of emotions beyond the basic positive, negative, and neutral can enrich the analysis. Time series analysis to observe how sentiments evolve over time is also valuable. Social network analysis can identify key influencers and community sentiment differences. Advanced machine learning techniques, like deep learning, can improve accuracy and reliability. Lastly, ethical considerations and privacy protection are paramount in data collection and analysis. By combining these approaches and improvements, a more thorough and accurate sentiment analysis of the Israel-Palestine conflict can be achieved.

The Israel-Palestine conflict has seen a significant shift in global public opinion, increasingly favoring Palestine, largely due to the widespread dissemination of information through social media. This shift challenges traditional narratives and exposes the complexities of the situation on the ground. Israel's military actions in Gaza, often defended by Western countries, have drawn criticism and impacted its international reputation, potentially straining relations with Western allies.[14] The evolving situation raises concerns about further violence and instability in the region, highlighting the urgent need for dialogue, diplomacy, and a comprehensive approach to peace. Recognizing the rights and aspirations of both Israelis and Palestinians is essential in navigating the complexities and seeking a balanced and lasting resolution to this enduring conflict.

6 References

[1] Gelvin J L. The Israel-Palestine conflict: One hundred years of war[M]. Cambridge University Press, 2014.

[2] Azam S. The Analysis of Palestine Conflict and UN Role[C]//Proceedings of the 22nd International RAIS Conference on Social Sciences and Humanities. Scientia Moralitas Research Institute, 2021: 224-231.

[3] Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association, Volume 18, Issue 5, September 2011, Pages 544–551.

[4] Kong, Y., Meng, F., Carterette, B.: A Topological Method for Comparing Document Semantics. arXiv preprint arXiv:2012.04203 (2020)

[5] Konieczny, J.: Training of neural machine translation model to apply terminology constraints for language with robust inflection. Annals of Computer Science and Information Systems 26, 233234 (2021)

[6] Hanslo, R.: Evaluation of Neural Network Transformer Models for Named-Entity Recognition on Low-Resourced Languages. In: 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 115–119. IEEE (2021)

[7] Carroll Doherty. Americans' Views of the Israel-Hamas War[M]. Pew Research Center, PP_2023.12.07

[8] De Francisci Morales, G., Monti, C. & Starnini, M. No echo in the chambers of political interactions on Reddit. Sci Rep 11, 2818 (2021).

[9] Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. Social Media + Society, 7(2).

[10] Long K., Vines J., Sutton S., Brooker P., Feltwell T., Kirman B., Barnett J., Lawson S. (2017). "Could you define that in bot terms"? Requesting, creating and using bots on Reddit. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 3488–3500). Association for Computing Machinery.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, (3/1/2003), 993–1022.

[12] Carroll Doherty. Americans' Views of the Israel-Hamas War[M]. Pew Research Center, PP_2023.12.07

[13] De Francisci Morales, G., Monti, C. & Starnini, M. No echo in the chambers of political interactions on Reddit. Sci Rep 11, 2818 (2021).

[14] Dajani Daoudi, Mohammed S., and Zeina M. Barakat. "Israelis and Palestinians: Contested Narratives." Israel Studies, vol. 18, no. 2, 2013, pp. 53–69. JSTOR.