Voices and Geopolitics: Exploring Public Opinions and Official Policy in the Israel-Palestine Conflict

Kshitij Makwana, Sandeep Jala, Ishan Saksena
School of Information, University of Michigan, Ann Arbor

Introduction

In response to the escalating global discourse surrounding the Israel-Palestine conflict, this project scrutinizes the disjunction between official policies and public sentiments, specifically within the realm of social media, with a focus on Reddit. Targeting key global subreddits, we extract and analyze posts and comments using keywords from diverse news outlets.

Our primary objectives include deriving insights into public opinion on the Israel-Palestine conflict, determining the optimal machine learning model for conflict-specific classification, and assessing the proficiency of the GPT-3.5 model in accurately classifying textual data. Through this analysis, we aim to offer valuable contributions to understanding public sentiment dynamics in a complex geopolitical issue. By studying social media platforms like Reddit, we can better understand how public opinion on the Israel-Palestine conflict is shaped and expressed. This analysis allows us to identify discussion patterns and trends, providing valuable insights into the dynamics of public sentiment surrounding this complex geopolitical issue. Ultimately, our goal is to contribute to a more nuanced understanding of the conflict and its impact on society.

Methodology

Figure 1 shows the overall project workflow and provides a high-level overview of the work done.

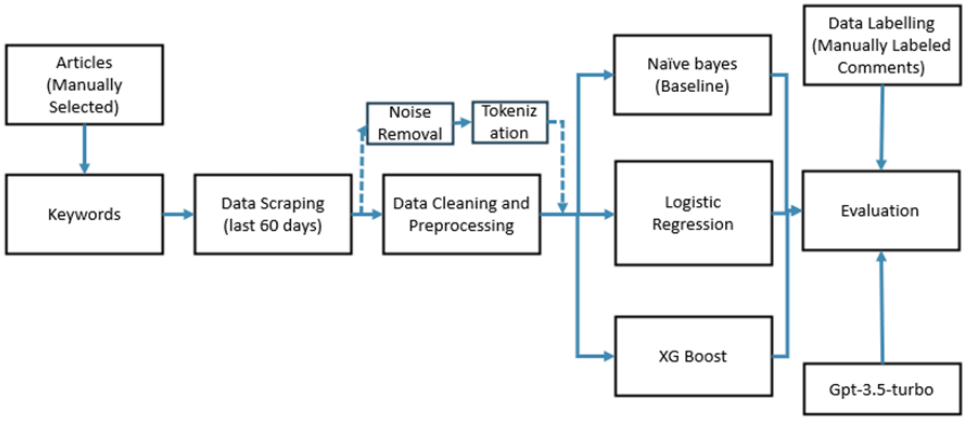

Figure 1: Workflow

Data collection

Our data collection had two significant aspects: keyword extraction and data scraping.

1. *Keyword Extraction* – We identified news articles from the top publishers concerning substantial events related to the conflict. From these articles, we used the newspaper3k [1] library in Python to extract keywords from these articles. The **newspaper** library is a Python library designed for web scraping and extracting information from online news articles. It simplifies fetching, parsing, and extracting content from news articles on the web.

2. *Data scraping* – We chose to scrape comments from Reddit using its API in Python. The keywords extracted in the first part were used as a search query to find posts related to the conflict. The search was done for posts from 1st October to 28th November. After the data was scraped, any irrelevant posts were discarded manually. The top 200 comments from each post were scraped. This was done to avoid Rate Limit errors and to maintain consistency across subreddits. The subreddits used were AskIndia, India, AskAnAmerican, Politics, Japan, Canada, AskCanada, MiddleEast, and Iran.

Preprocessing and data cleaning

The data cleaning steps considered for our use case are as follows –

1. *Removal of HTML tags* –We used **BeautifulSoup** to remove HTML tags from the text.
2. *Removal of user mentions and URLs* – Comments often include URLs and mentions of other users. We used regex to remove these.
3. *Removal of Punctuation and Special Characters*
4. *Conversion to lowercase*
5. *Lemmatization* - It is an NLP technique that reduces words to their base or root form, known as a lemma. The purpose of lemmatization is to normalize words, reducing them to a common base form. We used ‘spacy’ to perform lemmatization.
6. Removal of stop words - Stop words are common words that often carry little meaning and are frequently used in a language. By eliminating stop words, the emphasis is placed on content words that carry more semantic meaning.

These data cleaning steps ensure the text classifier is trained on a high-quality dataset, free from unnecessary noise and inconsistencies.

Table 1 summarizes the number of posts and comments scraped and cleaned.

Country	No. of subreddit	No. of posts	No. of comments	No. of comments labelled
Canada	2	20	4850	286
India	2	81	5225	300
Japan	1	14	978	231
Middle east	2	40	469	203
United states	2	22	1734	374

Table 1: The number of posts and comments scraped and cleaned.

Data Labeling

1. *Manual Labeling* – Comments were manually labeled considering the context of the post and the events that occurred during the current conflict. Comments in favor of Israel were labeled -1, in favor of Palestine were labeled 1, and those stating facts, being neutral or irrelevant to the topic were labeled 0. Table 2 gives examples of the comments and how they were labeled.

2. *GPT labeling* – GPT 3.5 Turbo 1106 was also used to generate labels for our data. The OpenAI API [2] was used for this purpose. API costs and tokens available for output generation were considered while selecting the GPT model. The performance of the GPT model is discussed later in the results section.

Sample Comment	Manual Label	GPT label
Palestine because Israel is a settler colonial State built on the blood of the native people. The Israelis are not native to the land, they are all European migrants.	1	0
I dont think we can afford to be involved in it.	0	1
Iran-REGIME (which was trained and armed and helped into power by the animal Arafat) not Iran. We the people of Iran know the truth and stand with Israel.	-1	0

Table 2: Comments labeled manually and with GPT

TF-IDF Vectorization - TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a numerical representation technique that converts a collection of documents, such as texts or articles, into a matrix of numerical features. The output is a numerical vector representing the document, where each element corresponds to the TF-IDF score of a specific term.

Model Selection and Training – We trained three models on our data and compared their performance with GPT's.

1. Naïve Bayes - Naive Bayes [3] is a simple probabilistic classification algorithm based on Bayes' theorem. The model is fundamental and serves as a good baseline. It makes some very strong assumptions about the independence of the data, which might not always be accurate. Multinomial Naïve Bayes was used as a baseline for our project as it naturally extends to multiple classes. It models the probability

distribution of the word features given the class, and the class with the highest probability is assigned to a document.

2. Logistic Regression - Logistic Regression [4] is a statistical model used for classification tasks. It models the probability that an instance belongs to a particular class using the logistic function. Logistic Regression was used for this task due to its high interpretability, efficiency, and robustness to noise commonly seen in textual data. For multi-class classification, such as ours, Logistic Regression defaults to the one-vs-rest (OvR) [5] scheme.

3. XGBoost Classifier - XGBoost (Extreme Gradient Boosting) [6] is a powerful and efficient machine learning algorithm that belongs to the class of gradient boosting algorithms. It minimizes a loss function by adding weak learners in a way that optimally reduces the overall prediction error. XGBoost was used for our task due to its efficiency, speed, and ability to handle high-dimensional and sparse feature spaces inherent in social media data.

Other models, including Deep neural networks with Bidirectional LSTMs and BERT, were tried for this task but failed to perform efficiently due to a small number of labeled comments.

Model Evaluation

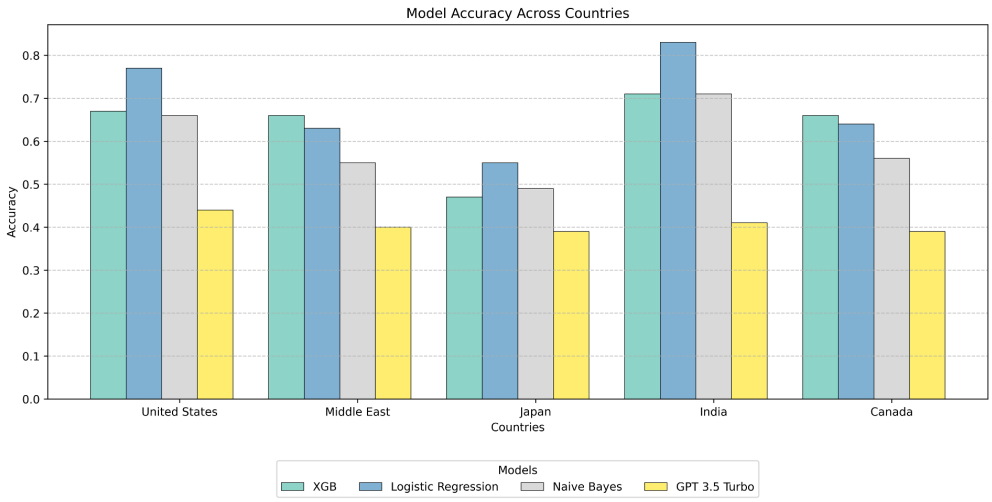

Figure 2: Accuracy of the models

Figure 2 shows our three text classification models and the accuracy of GPT 3.5 Turbo on the data across countries. The Naive Bayes was our baseline. We can see that our other models beat Naive Bayes in all cases. GPT 3.5 Turbo struggled to classify the data, achieving an average accuracy of 41%. These are the class accuracies for the XGBoost classifier on the India dataset in Figure 3.

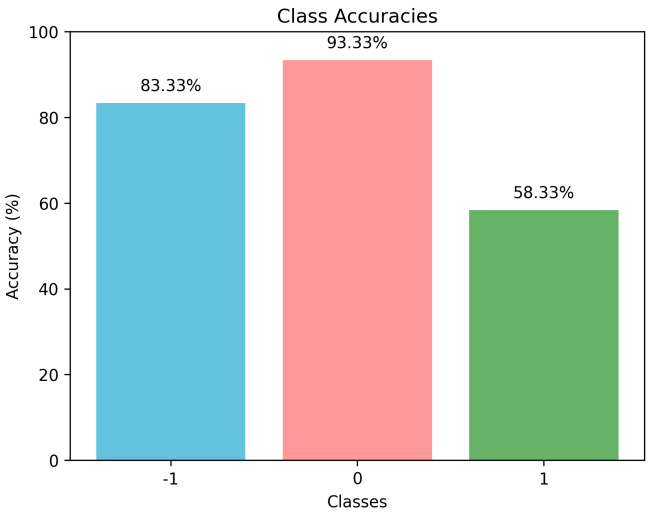

Figure 3: Class Accuracy

Analysis

Based on the results of the United Nations General Assembly resolution calling for a humanitarian truce between Israel and Hamas, we determined the official stance of the region as Palestine, Neutral, and Israel based on whether they voted For, Abstained from voting, or Against the resolution respectively. Table 3 shows the official stance, the percentage distribution of public opinion, and if the majority of public opinion matches the stance. [Middle East: Iran, Saudi Arabia, UAE, Bahrain, Turkiye]

Country	Official stance	Public Opinion (In percentage)			Govt. policy vs. Public opinion
		Neutral	Israel	Palestine	
Canada	Neutral	56.6	24.4	19	Match
India	Neutral	47	34	19	Match
Japan	Neutral	39	35	26	Match
Middle East	Palestine	57	36	7	Doesn’t Match
United States	Israel	62	28	10	Doesn’t Match

Table 3: Government policies of the countries vs. the public opinion

BERT Text summarization: We used Google's BERT [10] to corroborate our conclusions about the stances of various countries. As a transformer-based model, BERT (Bidirectional Encoder Representations from Transformers) employs a bidirectional approach to understand the context and relationships between words in a sentence. BERT captures the most salient information through its layered neural network architecture, distilling it into a coherent and concise summary.

BERT was provided with a concatenated string of all comments of a particular subreddit and told to summarize the contents. Then, the output of BERT was given to GPT-4 to interpret into an even shorter summary. This summary was used to confirm the stance of a region.

Example: GPT-4 summary for r/MiddleEast
The r/MiddleEast subreddit reflects diverse opinions on the Israeli-Palestinian conflict. Users discuss their stances on ceasefires, express skepticism about media objectivity, and share varied perspectives on the political dynamics between Israel and Hamas. Some users voice support for Israel, while others criticize media representations. The discussions underscore the complexity of viewpoints within the Middle East community on Reddit. While some comments express support for Israel, the majority appear neutral, reflecting a range of perspectives among participants in the subreddit.

Related work

The literature on sentiment analysis takes different approaches. As exemplified by [7], early approaches focused on syntax-based approaches, which were developed to evaluate verbs. Recent advances support nuanced sentiment analysis using rule-based machine learning techniques.

In a more specific case, [8] examines the prediction of politics in online discussions using natural language processing techniques. The study summarizes the results and suggests effective strategies for differentiating political positions in informal online discussions.

Turning to natural language processing models, [9] examines the evolution of GPT series models and highlights their performance over time. This is consistent with the objective of our project, which is to test the skill of GPT-3.5 in classifying textual content related to the Israeli-Palestinian conflict. The study provides valuable insight into the potential of state-of-the-art models.

Overall, this study lays the groundwork for our work and informs our investigation of emotional dynamics, political prediction, and the possibility of advanced natural language processing models in the context of social media conversations about the geopolitics of the countries.

Discussion

1) Challenges in Labeling:

a) Being a complex classification task, dealing with opinions with mixed neutral and supportive arguments was challenging.
b) Some opinions needed advanced human reasoning beyond simpler models' capabilities.

2) GPT's failure in Labeling: GPT 3.5 Turbo had an average of 41% accuracy in labeling the data. This unexpectedly suboptimal performance of GPT (even after heavy prompt engineering and zero-shot training) can have several reasons:

a) GPT's policy constraints on sensitive topics caused it to remain neutral or refuse to classify comments
b) GPT had difficulty in identifying emotions within a limited context. GPT 3.5 Turbo was trained till September 2021, years before the recent conflict (although the war is long-standing). It is possible that GPT could not process the context given to it by us via prompt engineering.
c) GPT 3.5 Turbo has limitations working with longer tasks.

Future Work

A. Smarter strategies for Data Labeling: To overcome the challenges faced while labeling data, we can work on smart strategies to handle these complexities. Semi-supervised learning was also considered but dropped due to lack of time.
B. Sarcasm detection and other features: The simple machine learning models faced problems predicting sarcasm, humor, or mockery comments. By incorporating more features like sarcasm, we hope to improve the quality of our predictions.
C. Refining data-scraping and cleaning processes: Many of the project's limitations were due to limited computing power and resources. With better computing power, we can increase the quality and quantity of data.
D. Temporal element: An interesting variable to consider in the context of our analysis would be time. It would be insightful to see how the world's opinions shifted throughout the events of the conflict. A temporal element would help understand the data better and improve machine learning models predicting stances.
E. Contextual understanding and local languages: Some comments in subreddits like r/India and r/AskIndia were in Hindi. Some comments were in English but required some Indian context to understand. Future work can be done on these differences to understand the data better and help machine learning models classify the data.

Conclusion

In the process of completing this project, we acquired practical experience in extracting and scraping data and learned essential data cleaning techniques. We gained hands-on experience in training and evaluating ML models and developed proficiency in conveying project insights through a poster presentation. Overall, this project equipped us with a holistic set of skills.

With our model demonstrating promising performance even with a limited dataset, there is potential to glean valuable insights into public opinion on social media. This preliminary success suggests that leveraging machine learning tools for sentiment analysis can inform the creation of policies that better align with the public. As we scale

our efforts with more extensive data and enhanced computing power, this serves as a hopeful proof of concept, showcasing the possibility of employing advanced ML techniques to generate more informed and responsive national policies considering public sentiment.

References

1. Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics (pp. 63–70). Association for Computational Linguistics. [GitHub Repository](https://github.com/codelucas/newspaper)
2. OpenAI. (2023). OpenAI GPT-3.5. Retrieved from [OpenAI Models Documentation](https://platform.openai.com/docs/models)
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830. [Scikit-learn Naive Bayes Documentation](https://scikit-learn.org/stable/modules/naive_bayes.html)
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830. [Scikit-learn Logistic Regression Documentation](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
5. Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. Journal of Machine Learning Research, 5, 101-141. [PDF](https://www.jmlr.org/papers/volume5/rifkin04a/rifkin04a.pdf)
6. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794). [XGBoost Documentation](https://xgboost.readthedocs.io/en/stable/python/index.html)
7. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. In Foundations and Trends® in Information Retrieval (Vol. 2, No. 1–2, pp. 1–135). Now Publishers Inc. [Link to Chapter](https://link.springer.com/chapter/10.1007/978-81-322-2250-7_46)
8. Liu, B. (2015). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 8(1), 1-167. [DOI: 10.2200/S00616ED1V01Y201501HLT025]
9. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2023). Language Models are Few-Shot Learners. arXiv preprint arXiv:2303.10420. [arXiv:2303.10420](https://arxiv.org/abs/2303.10420)
10. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. [PDF](https://arxiv.org/pdf/1810.04805.pdf)