"English-Speaking Twitter Users Display Negative Sentiment Towards Israel-Palestine Conflict"

Group 7: Quan Ly, Michelle Odnert, Kaiden Ong, Jonathan Osorio, Stephanie Wang

1. Abstract

This project employs sentiment analysis and machine learning models to gauge English-speaking Twitter users' sentiments toward the Israel-Palestine conflict. Utilizing the Octoparse API, tweets were scraped under hashtags such as #IsraelPalestineConflict, #ProIsrael, #ProPalestine, etc. from October 7, 2023 (the date of the Hamas attack), onwards. The final dataset comprises of 1,360 rows including polarity, date posted, user, tweet content, and likes. The analysis involves comparing Roberta and Vader polarity models and using machine learning (Decision Trees, Support Vector Machines, Naïve Bayes) to categorize tweets into pro-Israel, pro-Palestine, or neutral. From there, the performances were analyzed through metrics like accuracy, precision, and recall on the testing data. The Roberta model outperformed Vader, revealing predominantly negative sentiments. Machine learning models achieved an accuracy of ~53%, with insights indicating a majority of tweets leaning towards Pro-Palestine. Ethical considerations include privacy in data scraping and responsible result presentation. Future directions include model refinement and dataset expansion.

2. Introduction

In the age of social media, the daily data influx is a valuable resource for gauging public opinions on diverse topics. Motivated by applying in-class machine learning concepts to a current real-world issue, the project delves into the Israel-Palestine conflict. The focus is on capturing and analyzing public sentiment expressed by English-speaking Twitter users, recognizing the limitations of traditional methods such as opinion polls in providing nuanced insights. The project seeks to address this gap by offering a more accurate and effective alternative to traditional polls, leveraging sentiment analysis and machine learning models. The study used Twitter data from October 7, 2023, employing models such as Vader, Roberta, Support Vector Machines (SVM), Decision Tree (DT), and Naïve Bayes (NB). The findings reveal a prevailing pro-Palestine sentiment, highlighting the challenges of sentiment classification on Twitter. Through model refinement, we anticipate contributing to a more nuanced understanding of public sentiment and evaluating the potential impact.

3. Related Work

Other projects around this topic include sentiment analysis studies around other topics. The paper 'Sentiment Analysis on Twitter Data' conducted a general sentiment analysis study on Twitter data using machine learning algorithms. The study categorizes tweets into positive, negative, or neutral sentiments related to specific query terms, enabling applications to assess customer feedback for product improvement (Sahayak et al., 2015). Additionally, a project specifically analyzed Twitter tweets from December 2022 to January 2023 about ChatGPT (Huang, 2023). Three research questions were asked regarding the main topics and sentiments of the conversations around early ChatGPT users. The study allowed us to understand and assess ChatGPT's capability, effectiveness, and challenges. A third study that was found to be the most similar to our project was 'Sentiment Analysis of Political Tweets for Israel using Machine Learning' (Gangwar et al., 2022). The creators analyzed tweets in May 2021 and scraped data using hashtags #IsraelUnderAttack, #IStandWithIsrael, #WeStandWithIsrael, and #IsraelPalestineConflict. They utilized SVM, DT, and NB with the NB model having the highest accuracy of 93.21%. However, upon inspection, some errors in their code led to the accuracy being higher than expected.

4. Data

To better understand public opinion on the Israel Palestine situation, we decided to focus on Twitter sentiment by scraping and analyzing tweets as our data set. We believed that Twitter provides real time opinions on current news, and since this is a recently popular topic there would be highly active opinions on the situation. In total we scraped 1,360 tweets. We found that the highest sentiment count was Pro Palestine at 600, followed by Neutral at 410, then Pro Israel at 350. For hashtags, #ProPalestine came up to be 451, #ProIsrael at 397, and #IsraelPalestineWar at 231.

5. Approach

The methodological framework for this project involved a systematic four-step process: overall sentiment analysis, feature engineering, training/testing, and the comparison of model metrics. In the initial stage, natural language processing was applied to determine the collective sentiment of Twitter users. This analysis involved computing the score of each tweet based on the hashtag by which it was categorized. The Vader model assessed the compound score, while the RoBERTa model provided a breakdown of sentiment categories for each hashtag-associated tweet.

Moving to the second stage, team members manually classified tweets into Pro-Israel, neutral, or Pro-Palestine categories, creating a new "polarity" feature. In the third stage, Decision Trees, Support Vector Machines (SVMs), and Naive Bayes were employed to train and test the models using this new feature. Finally, accuracy and precision scores were compared across the machine learning models to determine the best-performing one.

6. Experiments

The techniques used for sentiment analysis are as follows:

6.1 Vader Model

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a natural language processing model used for sentiment analysis. It assigns sentiment scores (positive, negative, or neutral) to words and phrases, considering context and grammar rules. Vader utilizes a sentiment lexicon and a set of grammatical rules to capture sentiments, making it suitable for analyzing the emotional tone of the text in various contexts. (Amanmyrat Abdullayev, 2022)

6.2 Roberta Model

RoBERTa (Robustly optimized BERT approach) is another natural language processing model used for semantic analysis. Roberta excels in capturing contextualized representations of words due to it being trained on a Twitter dataset, making it highly effective for tasks like sentiment analysis and text classification. Its training approach including an increased attention to textual context contributes to a more advanced performance compared to the Vader model. The Roberta model similarly provides scores in positive, negative, and neutral categories. (Amanmyrat Abdullayev, 2022)

6.3 Support Vector Classifiers

Support Vector Classifiers are supervised machine learning algorithms designed for classification and regression. Their goal is to find a hyperplane in a high-dimensional space that maximizes the margin between different classes of data points. Support vectors, the data points closest to the hyperplane, play a crucial role in determining the optimal separation. Once the optimal hyperplane is identified, SVM can efficiently classify new data points by determining on which side of the hyperplane they fall. SVMs are effective in moderate-sized datasets, offering robustness against overfitting, but can become computationally expensive. (*1.4. Support Vector Machines*, 2023)

6.4 Decision Trees

Decision Trees are a machine learning algorithm also used for both classification and regression. They have a flowchart-like structure where each internal node denotes a decision based on a feature, each branch signifies the outcome of that decision, and each leaf node represents the final predicted label or value. The goal is to recursively split the data into subsets based on the most informative features, creating a hierarchical set of decision rules. (*1.10. Decision Trees*, 2023)

6.5 Naive Bayes

Naive Bayes is a probabilistic classification algorithm. It is based on Bayes' theorem, which calculates the probability of a hypothesis given observed evidence. The "naive" assumption in Naive Bayes is that features used for classification are conditionally independent, meaning the presence or absence of one feature does not affect the presence or absence of another. The algorithm calculates the probability of each class given the input features and selects the class with the highest probability as the predicted class for a given instance.

7. Results

Before looking directly at the results of running the overall sentiment analysis experiment, it’s important to understand the fundamental differences between the VADER and RoBERTa models.

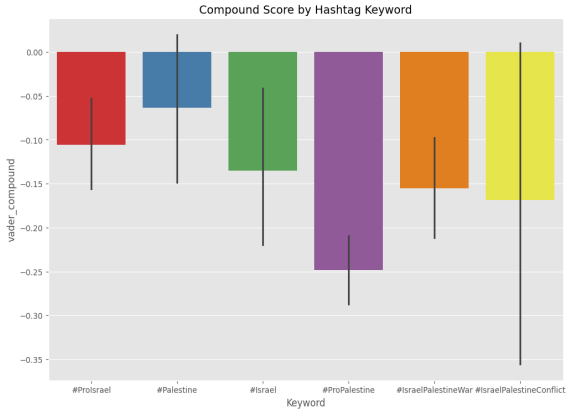
7.1 VADER

The VADER model takes a lexicon based approached, breaking down Tweets word by word to give the overall content a score between -1 and 1. This approach works well to capture basic sentiments across a wide range of topics; however, it can struggle to discern more nuanced and context based sentiments. Taking a look at Figure 1 below, the VADER model categorizes several hashtags such as #Palestine and #IsraelPalestineConflict as slightly negative or even neutral, when taking into account the margin of error.

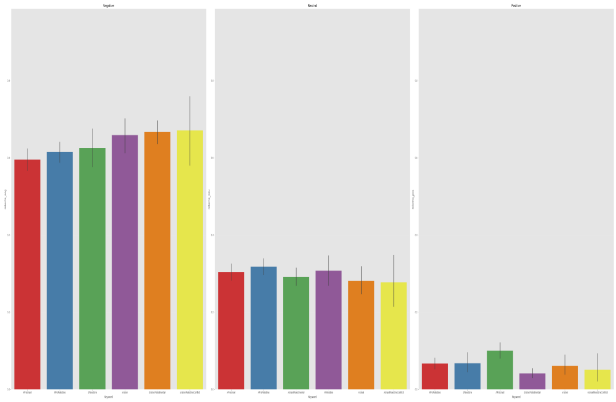
7.2 RoBERTa

On the other hand, the RoBERTa model was trained using deep learning techniques, thus is able to analyze context and understand human semantics at a deeper level. This resulted in the RoBERTa model finding many more tweets to have a negative sentiment. Now looking at, Figure 2, the RoBERTa model clearly identifies a majority of the tweets as negative and very few as negative.

VADER Model (Figure 1):



RoBERTa Model (Figure 2):



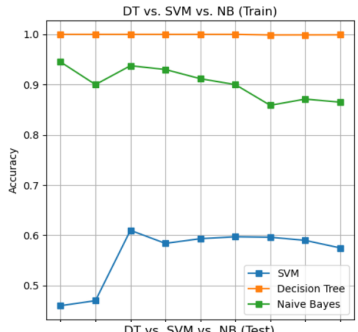
VADER vs. RoBERTa Sentiment Distribuation (Figure 3):

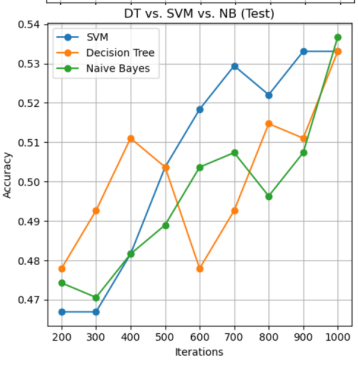
Negative	Neutral	Positive
~[0.09 - 0.13] (V) **~[0.59 - 0.67] (R)**	**~[0.79 - 0.84] (V)** ~[0.27 - 0.32] (R)	~[0.05 - 0.09] (V) ~[0.04 - 0.10] (R)

The results of the two models gives two different answers to this experiment. The VADER model shows that approximately 80% of the data has neutral sentiment, while positive and negative both take up about 10% of the data. On the other hand, the more complex RoBERTa model demonstrates a strong negative sentiment towards the Israel Palestine conflict, categorizing about two thirds of the tweets as negative. In summary, the results underline the RoBERTa model's superior capability in discerning nuanced negative sentiments. However, regardless of the biases and opinions that may form within each hashtag on twitter, both models suggest that the overall sentiment on the Israel Palestine conflict is negative.

7.3 Training ML Models

The training and testing results of Decision Trees (DT), Support Vector Machines (SVM), and Naive Bayes (NB) reveal intriguing patterns in their performance metrics. During training, the DT model had nearly perfect accuracy in training at 99%, with NB following closely behind at around 90%. However with SVMs, there was a significant drop off, all the way down to approximately 60%. However, the important part is the testing accuracy, and between these three models, NB had the best accuracy, but just barely at 56%, followed by DT at 54%, and SVM at 53%. This convergence of test accuracies show that even though DT and NB had significantly higher training accuracies, SVM was the one that actually generalized the best and overfitted the least.





Moreover, observing the overarching trend of testing accuracies, it is clear that as the iterations increase so does the accuracy. In this context, iterations represents how much of the training data is used (e.g., 200 representing a subset of 200 tweets used in training). Since the testing accuracies are still increasing this means that increasing training data by scraping more tweets would likely result in higher accuracies, and possibly a larger discrepancy between the three models.

While NB did achieve the best testing accuracy, it is not clear whether it's the "best model". However, taking into account related works and other similar research, this result is supported, making it a valid assumption and conclusion.

8. Discussion

The findings of this study unveil intriguing aspects regarding the performance of all five sentiment analysis models VADER, RoBERTa, Decision Trees, Support Vector Machines, and Naive Bayes. Notably, the Vader and RoBERTa models both revealed a prevailing negative sentiment associated with the Israel-Palestine conflict, regardless of sides and opinions. RoBERTa, with its deeper contextual understanding, had superior performance in capturing nuanced negative sentiments, while VADER disregarded context and classified a majority of the data as neutral sentiment. Within the machine learning models, the discrepancy between training and testing accuracies across DT, SVM, and NB underscores the significance of generalization and overfitting. While DT and NB had higher training accuracies, their testing accuracies were outperformed by SVM, suggesting its ability to generalize better and avoid overfitting.

Overall, the contributions of this study lie in providing insights into the nuanced sentiment dynamics surrounding a complex socio-political issue. It highlights the capabilities and limitations of various sentiment analysis models, emphasizes the importance of context, data diversity, and sample size in model performance, and contributes to the ongoing discussion on sentiment analysis methodologies. Additionally, the trained and tested ML models provide a baseline for understanding the highest performing models in analysizing sentiment within the context of the Israel Palestine conflict, allowing for future classification of non-catergorized tweets and data. These insights lay a foundation for further research in refining sentiment analysis techniques for understanding these intricate political biases and opinions.

9. Limitations and Future Work

Given the scale of the project, there were quite a few limitations. First, with only 1,360 tweets, we did not obtain anything close to a representative dataset - these were simply the most popular tweets from our time frame. The decision to have such a small dataset was due to our approach of classifying the Tweets by hand - for this, around 1400 tweets was a reasonable amount. The second limitation - by splitting up the dataset for each team member to help with classification, there were inconsistencies in labeling that may have played a role in our low accuracy. This comes into play for the third limitation - some tweets do not fit well into one of the categories of pro-Israel, pro-Palestine, or neutral. For example, someone who expresses sentiment against Hamas is not necessarily in favor of Israel. In future work, we want to expand our classification categories to better accommodate these nuances. Additionally, we would like to scrape more tweets to gain a more representative sample and could use an API that includes demographic information (with all due privacy considerations). Finally, there is further feature engineering we can do to prepare the text for the models, and include media (photos/videos) in the tweets as part of the features.

10. Ethical Considerations

Analyzing Twitter sentiment along with the Israel-Palestine conflict brings up many ethical considerations including user privacy, algorithmic bias, and emotional impact. Obtaining informed consent to use real people's tweets for research should be recognized. Avoiding oversimplification in using diverse categorization should be approached in order to respect different peoples' opinions. Finally, handling the data with care must be made aware of the potential emotional impact from the content itself.

11. Conclusion

This project revealed several significant insights. Firstly, regardless of the distribution, an overarching trend emerged: the tweets consistently conveyed a negative sentiment. This may just be because conflict and war are dismal subject matters. However, this observation could prompt discussions about the current state of social media, raising questions about whether individuals or the system itself may be pushing certain ideas or beliefs. Moreover, the training and testing phases unveiled machine learning models' susceptibility to overfitting. Performance metrics exhibited a decline ranging from 10% to nearly 50% between the training and testing stages. However, it is important to state that the Naive Bayes model outperformed others with an accuracy of ~54%. Despite initial impressions, this score significantly surpasses the randomness baseline for determining text polarity. Given the project's structure, tweets were categorized into three classes. Therefore, a random selection would yield an accuracy of ~33%, making the Naive Bayes model ~21% more accurate than chance, a noteworthy accomplishment in this context.

12. References

1.10. Decision Trees. (2023). Scikit-Learn. https://scikit-learn.org/stable/modules/tree.html#tree

1.40. Support Vector Machines. (2023). Scikit-Learn. https://scikit-learn.org/stable/modules/svm.html

1.9. Naive Bayes. (2023). Scikit-Learn. https://scikit-learn.org/stable/modules/naive_bayes.html

Amanmyrat Abdullayev. (2022, October 14). *Sentiment Analysis with VADER and Twitter-roBERTa - Amanmyrat Abdullayev - Medium*. Medium; Medium. https://medium.com/@amanabdulla296/sentiment-analysis-with-vader-and-twitter-roberta-2ede7fb78909

Gangwar, A., & Mehta, T. (2022). Sentiment Analysis of Political Tweets for Israel using Machine Learning.

Huang, H. (2023, May). Twitter Sentiment Analysis about ChatGPT. GitHub. https://github.com/hxycorn/Twitter-Sentiment-Analysis-about-ChatGPT

Rajasree, R., & Neethu, M. S. (2013). Sentiment Analysis in Twitter using Machine Learning techniques. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*.

Sahayak, V., Shete, V., & Pathan, A. (2015, January). Sentiment Analysis on Twitter Data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *2*(1).