אוניברסיטת
אריאל
בשומרון

SENTIMENT ANALYSIS USING MACHINE LEARNING ON THE NEW YORK TIMES HEADLINES

Industrial engineering and management

Advanced topics in machine learning

Dr. Chen Hagaj

Meital Padwa

Bar Mizrahi

Table of contents

Abstract

This project is about The Yew York Times reporters' tendencies regarding the October war occurring in the Israel since October 7th,23.
The purpose of this project is to analyze the sentiment of the NY Times war - related articles.
Machine learning sentiment analysis methods were used to see whether the writers' opinions are more Pro-Israel or Pro-Hamas. In addition, topic analysis was implemented to review the main topics that emerge from the articles.

The results show consecutive sharp drop in the polarity score to negative values for Israel- related articles since the war began while showing an incline in the number of articles published concerning Israel.
Despite the results, the conclusion drawn from the above is that there is no clear stance taken by the NY Times when drilling down to the articles with most negative polarity score values.

Key words

NY Times, Sentiment Analysis, Israel, Topic Analysis, VADER, Tokenization.

Introduction

During the recent ongoing war between Israel and Hamas, there was an extensive backlash from communities around the world, including the U.S citizens.
The public opinion is known to be a key component in the decision-making process of the democratic governments and organizations around the world and affect directly on their actions.
The NY Times is a popular and a highly valued newspaper which is read and mentioned globally. Due to its large reach and impact, it very important to understand its stance on important topics.

The main question that comes up is if it is possible to determine the tendency of newspapers on Israel by using Machine learning models and methodologies.
The objective of this project is to examine the sentiment that arises from the articles related to Israel during the recent war and see if the NY Times is a non-bias new paper.

With the results of the project, it will be possible to determine whether Sentiment analysis using ML models can find the true opinion of the new and media and, perhaps, in the future, the public's opinion.

Dataset and Features

In this project the NYT API is used.
The API call is done with the assistance of " pynytimes" importing "NYTAPI", a library designated for NYT API calls.

The query was destined to pull all the possible articles from September 7th until January 7th of 2024.

The query contained only articles that have the word "Israel" in their header. The query specifies the types of materials required (only NYT articles and news), only articles that were published by the NYT (source), and that the articles bodies must contain the words " Gaza", "Hamas", and "Israel". Results are sorted from the oldest to the newest.

The initial dataset contains 2 columns and 615 rows. The columns are the headlines of the articles and their dates.
The next step was to clean the data so it will be easier to analyze. A tokenization process was used on the column 'headline' to remove all punctuation and converting all the text and letters to lowercases. The following step was removing all stop words by using the 'nltk.corpus' library, in addition to that a choice has been made to remove the following word as well: 'israel', 'israeli', 'Israeli', 'israelhamas', 'hamas', 'gaza', 'U', 'u' due to the fact that the frequency of their appearance in the text is extremely high and it is known to be so because of the specific query that is being asked. By removing these words, we make room for the other words to impact on the different model being used down the line. The last step was preforming 'WordNetLemmatizer'. 'WordNetLemmatizer' organizes words into sets of synonyms and records their semantic relationships. For example, for the word 'running' the returning word will be 'run'. The purpose of using this library is to normalize the data to the base and root of each word' as well as to improve text analysis in the following steps by ensuring that words with the same meaning are treated consistently, regardless of their inflected forms.

Methodology

Analyzing Article Quantity Over Time:

The quantity of articles published over time is analyzed to understand trends in article publication. Using a Grouping method, the number of articles per week is shown from the month before the war began and until the week of January 7th.

Creating Wordcloud of Cleaned Text:

A wordcloud visualization is created to illustrate the frequency of words in the cleaned text. The cleaned text from all headlines is concatenated to form a single string. Using the WordCloud library, a wordcloud is generated with a specified shape mask for design (the hostage ribbon), providing a visually representation of the most prominent words in the dataset.

Sentiment Analysis Using VADER:

The polarity scores are computed for both raw and cleaned versions of the headlines to see whether it makes a difference in polarity score, due to the conflicted opinions of should the data be cleaned when preforming SentimentIntensityAnalyzer or not.

Grouping and Aggregating Polarity Scores:

After computing the polarity scores, the dataset is grouped by date to analyze sentiment trends over time. For each date, the average polarity scores for both raw and cleaned data are calculated. This aggregation process provides insight into the overall sentiment on different days.

Visualizing Polarity Scores Over Time:

To visualize the sentiment trends, scatterplots of polarity scores over time are plotted for both raw and cleaned data. These scatterplots display the distribution of polarity scores for each headline, with a trend line overlaid to highlight the general sentiment trajectory over time.

Filtering Data Based on Polarity Scores:

The dataset is filtered based on specific criteria related to polarity scores. Dates where the average polarity score exceeds a predefined threshold (bigger than 0.1 which showcase positive sentiment) are identified. Additionally, headlines before a certain date (e.g., October 7th, 2023) with polarity scores above the threshold are included in the filtered data. This filtering process isolates headlines with significant sentiment polarity for further analysis.

Identifying Dates with Extreme Sentiment:

Dates with the highest and lowest average polarity scores are determined to pinpoint instances of extreme sentiment. This process was done in order to determine whether sentiment analysis is a correct way to determine a writer's stance or not.

Topic Modeling Using LDA:

Topic modeling is employed to uncover underlying themes present in the dataset. The topics are extracted and visualized using pyLDAvis, allowing for a deeper understanding of the main themes discussed in the headlines. The number of topics selected is 3 with a random.seed(4) because when selecting these two values combined a clear division of three different topic is very visible.

This detailed methodology outlines the step-by-step process on the provided code. It provides a structured approach to comprehensively analyze sentiment and topics within the dataset.

Experiments

API call

The " pynytimes" library was used to pull data from NYT website.
Start and end date were specified, and the options settings were arranged in a way that will return only NYT articles related to Israel (headline must contain Israel).

Preprocessing-

During the data cleaning and tokenization, a lemmatization process was applied for standardization of the textual terms. The Porter stemmer was not used due to its insensitivity to the context. Porter's stemming can return words that don't exist, unlike Lemmatizing which retains the term context.

Article trends

The cleaned text and date columns are extracted from the dataset, and the date column is converted to datetime format and set as the index. The data is then resampled to group articles by week, and the mean polarity score is calculated for each week. A line plot is generated to visualize the weekly article quantity over time, providing insights into fluctuations in publishing activity.

Sentiment analysis

Polarity score - The SentimentIntensityAnalyzer from NLTK's VADER is employed to gauge the sentiment of headlines in the dataset. This analyzer assigns polarity scores to each headline, which indicate the positivity, negativity, or neutrality of the text.

Article drill down - The dataset is then filtered to include headlines from the dates that had the extremist average polarity scores, allowing for a deeper examination of sentiment fluctuations and potential factors contributing to extreme sentiment. Furthermore, the headline of the specific articles was showcased as well as a manual search in the web was done to detect what happened according to different news sites in these specific dates.

Topic analysis

A Dictionary and Corpus are created for topic modeling. An LDA model is trained on the corpus to identify topics within the text.

Results

The sentiment score has declined significantly since Oct 7th, and almost always was in the negative values (figure 1).
The three main topics found with the topic analysis (figure 2):

- Topic 1- Represents the war .
- Topic 2- Represents more political and aid semantics.
- Topic 3- Represents a hostages' view.

In addition, the interest in the ongoing war was at first inclining and then a moderate decrease took place (figure 3).

Discussion

During the modeling of the sentiment on the article headers, there is no visible political side taken by the NYT. It was determined despite the results due to further investigation of the headers with the most negative values. The headers were not pro-Israel nor pro-gaza. The sentiment analysis considers some words more negative and does not take in account the context of the sentence. Hence, we cannot claim that the NYT are more Pro-Israel nor Pro-Hamas.
Additionally, there was no clear difference between the cleaned and raw data in the VADER polarity score model results. That being said, there is no preferred approach over the other (figure 1).
Finally, it is safe to say that the topic analysis showed good performance upon its results, splitting the data to three separate and logical topicts, War, Politics and Hostages.

Conclusions

This project was about The Yew York Times reporters' tendencies regarding the October war occurring in the Israel since October 7th,23. After implementing different models on the data, and comparing to the actual headlines it is noticeable that there is no clear stance coming from the writers despite the polarity scores values. This indicates that the current machine learning models are still lacking when it comes to accurate sentiment analysis. Furthermore, it is unclear whether stop words removal and lemmatization is necessary when it comes to preforming polarity score analysis when using the VADER engine.

Understanding the limitations of today's sentiment analysis when it comes to articles related to political and governmental issues can contribute to the future modeling development of opinions analysis. due to the possible contribution of this analysis, developing a proper model that will showcase the actual tendencies will be an optional future research.

Section 7- Contributions

Report- Bar

Presentation and GitHub- Meital and Bar

Code –

- Sentiment analysis, Wordcloud and preprocessing- Meital
- API call and data pull, Topic analysis, pronunciation, and explanations- Bar

Section 8- Appendices

Figure 1: sentiment analysis – polarity score as a function of time.

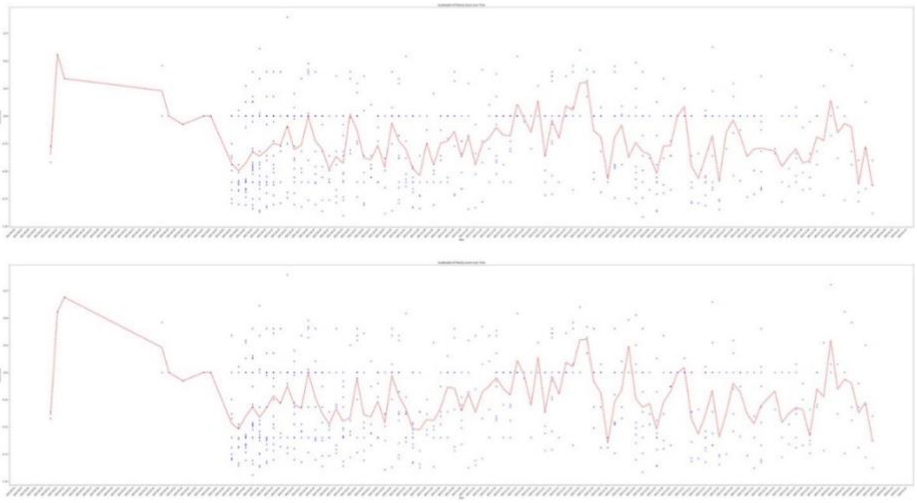


Figure 2: topic analysis

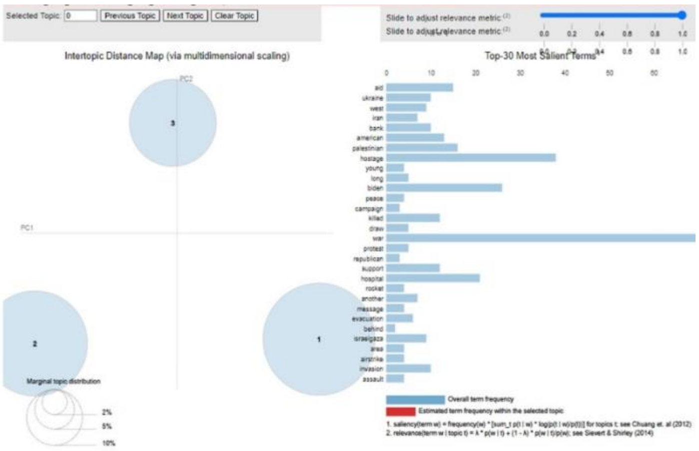
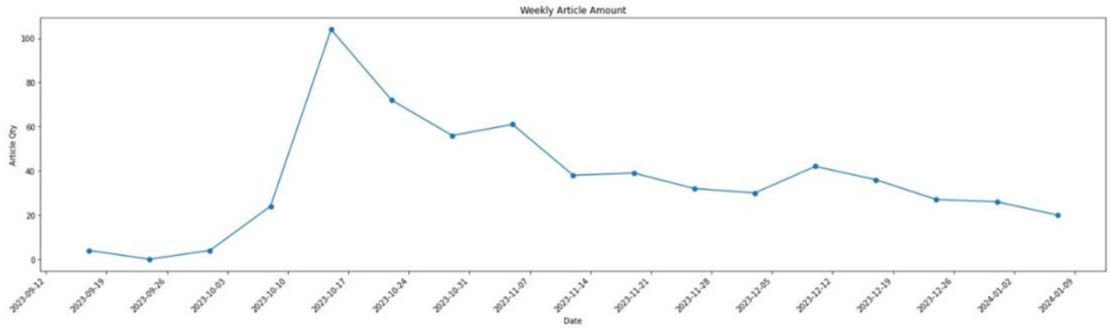

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms
0
10
20
30
40
50
60
aid
ukraine
west
iran
bank
american
palestinian
hostage
young
long
biden
peace
campaign
killed
draw
war
protest
republican
support
hospital
rocket
another
message
evacuation
behind
israelgaza
area
airstrike
invasion
assault
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3:



Similar related research:

V. Shirsat, R. Jagdale, K. Shende, S. N. Deshmukh and S. Kawale, &quot; Sentence Level Sentiment Analysis from News Articles and Blogs using Machine Learning Techniques &quot; International Journal of Computer Sciences and Engineering, vol. 7, no.5, pp. 1-6, May. 2019.