The Focus

On October 7th, 2023 violence of a scale never seen before in the history of Israel erupted near the border with Lebanon. Over 1,400 lost their lives in a series of brutal attacks in which the terrorist group Hamas specifically targeted civilians. Israel's response was relatively swift as it amassed tanks at the border and launched a ground invasion, all while shelling neighboring Palestine in a series of attacks that, while ostensibly targeting Hamas, also inflicted thousands of casualties on the Palestinian side.

For the media, covering these unfortunate events is a daunting task. Maintaining objectivity is challenging, given the suffering on both sides and the bias that certain news outlets have to one side–either Palestine or Israel. But how strong are these biases? And how do they change over time, especially when events on the ground are happening so rapidly? Understanding these biases is important because the media can shape both public opinion and public policy.

In this project, I will explore this question by specifically looking at the Op-Ed sections of three prominent media outlets–NYTimes.com, CNN.com, WSJ.com–which represent different parts of the political spectrum. My goal was to identify what these biases looked like, in terms of types of articles published, the language and tone employed, and the diversity of articles (do any of the three outlets try to portray both sides?). I tried to identify whether these biases shift and if so, to what extent might these shifts be related to what is happening in real-time versus some other factors. Ultimately, this rapidly evolving conflict can provide a glimpse into how sympathy in media coverage towards one side or the other happens in real time.

The scope of this project will include the methodology employed, the limitations encountered throughout, and an analysis, which includes NLP techniques and Generative AI. Finally, it will expand upon the ethical implications involved.

Methodology

Defining bias

Before I could analyze media bias around the escalating conflict, I needed to have a clearer notion of what I meant by "bias." My hope was to use NLP techniques that could do a political sentiment analyzer and that score would represent that "bias." Typically, that would mean to what extent does an article lean to the left or the right of the political spectrum. Here, the focus would be to what extent is an article supportive of Israel and to what extent is an article supportive of Palestine. Whether the political sentiment analyzer reflected this was moot: in the end, due to the excessive processing involved, I could not use models like BERT. (I will discuss this more in the limitations section.)

As an alternative, I decided to devise my own criteria for what I call a "sympathy score": to what extent is an article sympathetic towards the suffering inflicted on Israel and to what extent is the article sympathetic towards the suffering inflicted on Palestine.

The use of Generative AI

Simply coming up with a scoring system and then scoring it myself would not have been scalable nor would it have been a data science project. Therefore, I explored whether using Generative AI could help me devise a scoring system, but more importantly, score the articles quickly, I utilized two LLMs (large language models)--OpenAI's ChatGPT 4 and Anthropic's Claude–to score articles based on criteria I had established.

First, I had to translate these criteria into a prompt and input it into each model. The prompt itself invited some feedback on the "sympathy score" from the LLMs. Additionally, I provided them with some "benchmark" articles that I had already scored to establish reference points.

Most importantly, I also used the LLMs to help me refine my own criteria. This feedback was invaluable and also highlighted the inherent limitations and subjectivity involved in a project like this (see "Limitations" sections for more).

ChatGPT example:

You

Come up with a number between -1 and 1, which captures the extent to which an article sympathizes with Israel and to what extent it sympathizes with Palestine. To explain what I mean by "sympathize": the article I will provide in the next input is a -0.8 (strongly sympathetic to Israel). It focus almost exclusively on Israel's suffering. It will only mention Palestine in the context of Hamas, which is a internationally regarded terrorist organization and not the people of Palestine who are controlled and often oppressed by Hamas.

On the other hand, an article that is a positive 0.8 (included below the article that is a -0.8) will focus only on the current death and destruction in Gaza, painting Israel as the aggressor who is prepared to kill every last Palestinian to root out Hamas. It will not mention the horrific Oct. 7th acts on Israel or differentiate between Israel's government and its people.

Before I post the articles, let me know what I might have missed, assumptions that I should be aware of, and anything that is unclear.

1. **Subjectivity in Interpretation**: The biggest challenge with this approach is the inherent subjectivity. Different readers might interpret the same article differently based on their own biases and perspectives. Establishing clear criteria for what constitutes sympathy towards either side will be crucial.
2. **Complexity of Conflict**: The Israel-Palestine conflict is highly complex, with a long history and many nuances. Articles might contain elements that are sympathetic to both sides, or they might present a more nuanced view that doesn't clearly align with one side or the other. This complexity might be difficult to capture with a single number.
3. **Differentiating Between Criticism and Sympathy**: Criticism of one side's actions does not necessarily equate to sympathy for the other side. An article might be critical of Israel's military actions without necessarily expressing sympathy for Palestine, or vice versa.

I chose to use two LLMs vs. only one LLM, to be able to monitor for discrepancies and calibrate where necessary. This calibration involved "training" the model by providing feedback around its scoring, in those instances where the scoring was inaccurate. This took place for the first five articles, not including the "benchmark" articles. The process is captured below.

- I read the article and gave it a "sympathy score" on a scale from -1.0 (pro-Israel to 1.0 (pro-Palestine)
  - 0 was neutral, or at least not favoring one side over the other.
  - I did this before looking at the scores offered by the LLMs
- I dealt with discrepancies by providing feedback. (Claude was especially appreciative of such feedback.)
  - If one LLM provided a similar score as me, I would copy and paste its analysis into the other LLM for training (this happened throughout the project, when there was a discrepancy.)
- Sometimes I calibrated my own score based on the LLMs
  - When the two LLMs were within 0.1 of each other, I would revisit my rating, reading their analysis and asking follow up questions.
- For all the scoring (124 articles in total), if the models differed by more than .25, I would adjudicate, offering a final score.
  - When the discrepancy was 0.25 or less, I took the average of the two LLMs and did not read or score the article.
  - I ended up reading and scoring between 10-15% of the articles

<u>Choosing news sources</u>

An overarching question I wanted to explore is how media biases across the political spectrum differ. In order to avoid having my own assessment and subjectivity play a role, I sought out a third party (AllSides.com) to establish where a source fell on the political spectrum (see schematic below.)

Another criteria was that the news source published Op-Eds. Those publications marked as ‘C’, or politically in the center, did not have Op-Eds.

Additionally, the publication needed to have “enough” Op-Eds on the conflict so as to make the data analysis as robust as possible. Since this project took place during week 4 and 5 of the conflict, I therefore had to choose news sources that generated sufficient Op-Ed content.

Three criteria for media resources:

1. They had to come from across the political spectrum
2. They had to contain an Op-Ed section
3. They had to have a sufficient number of Op-Eds for the five week period

The three publications that best met these criteria were the New York Times (NYTimes.com), CNN (cnn.com), and the Wall Street Journal (wsj.com). In addition to meeting the three criteria above, these publications also have the distinction of being widely read and engaged with:

Source	Domain Authority

NYTimes.com	95
WSJ.com	95
CNN.com	96

(Domain authority is a Google ranking term that essentially shows how high a web page ranks. The scale is 0-100 putting these three sources at the high end for readership.)

Again, all articles were taken from the Op-Ed sections of these three news sources.

Op-Ed focus

Why Op-Eds? Op-Eds have a high reader engagement. NYTimes Op-Eds, for instance, can often elicit more than one thousand comments, which does not include the lively debate that emerges when readers comment on other comments. Furthermore, the Op-Ed section, which is not beholden to the standards of objectivity as the reporting sections. This can reveal a publication's editorial slant—which perspectives are included, which omitted, and whether there's a prevalent tone that skews one way or the other.

Finally, the tone of Op-Eds is stronger, allowing for a richer textual analysis, as well as a larger scope of sentiments.

The Dataset

In all, there were 58 Op-Ed pieces from the NYTimes.com, 21 from, and 45 from the WSJ. In addition to links to the article, the text to each article was also provided to go use basic sentiment analysis.

Here, I employed Textblob and was able to derive both Polarity and Subjectivity scores. I generated these first and identified that they were not consistent with the political sentiments and biases in the article. Coupled with the difficulty running more advanced NLP techniques on my laptop, this insight gave rise to the approach using Generative AI, detailed above.

Limitations

Ethical considerations

First and foremost, this is a very sensitive topic, one that has led to extreme anger and heartbreak across the political spectrum. It has exacerbated polarization in a country already strongly polarized. My intention here is not to weigh in on the political aspects of the conflict, and all the complexity that entails. Rather, I hope to explore how media sentiment differs across the political spectrum and how this sentiment can change over time, even in the same news source.

That said, this project is inherently infused with my own subjectivity. In other words, every aspect of the project relies on how I have decided to establish criteria around creating a "sympathy score" and identifying media biases around the Israel-Palestine conflict. Were another informed layperson (like myself) or a renowned expert to create their own rubric for "sympathy score", the results might be very different from those contained in this project.

Therefore, to clearly lay out my criteria in which I trained the LLMs:

- I focused on human suffering in both Israel and in Palestine.
- I took into account how much an article acknowledges the suffering on the "other side"
- I looked at how much a writer differentiates between a government and its people (Israel) and a terrorist organization and those under its thrall (Palestine)
- I included the tone and the intensity of that tone–e.g., harsh and uncompromising

Identifying these criteria is meant to communicate the dimensions of my subjectivity, not to imply that I was somehow able to somehow rise above my subjectivity.

Finally, given the immense human suffering involved, there is something unpalatable about the attempt to quantify this suffering. In fact, I questioned a few times whether or not to choose this topic. Nonetheless, being able to tease apart biases that exist in major media outlets, publications that shape our discourse, felt important regarding this conflict, given the perception that the U.S. media has historically taken sides in this conflict. Exploring these seemed like a worthwhile endeavor, possibly shedding a light on these biases and allowing for more equitable media coverage.

<u>My own limited knowledge</u>

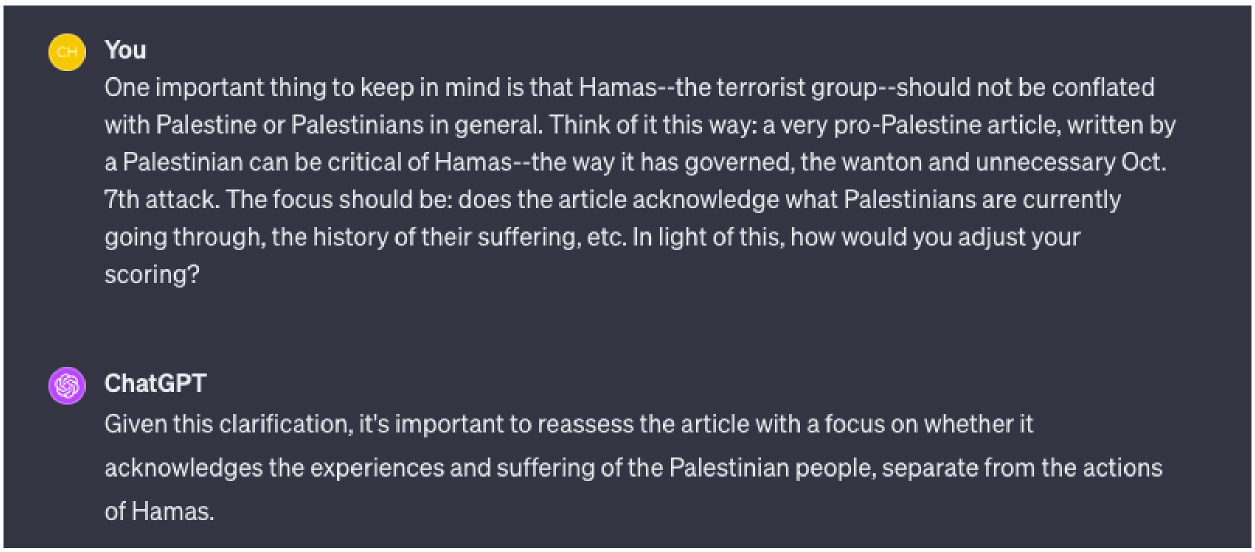
I am by no means an expert in this field. As an informed layperson, I have learned about the history of the region, at best, in broad strokes, which means I am unequipped to pick up on any over simplifications, where a writer might gloss over the complexities and nuances involved, or even outright falsehoods.

My hope is that the high standards of the media sources I selected and the scrutiny they are subjected to make it less likely that outright lies would flourish on their pages. Even when an article oversimplifies a complexity, which seems inevitable, my objective is to evaluate the political sentiment of an article, not judge how closely it adheres to what a panel of experts might deem "factual."

<u>The use of LLMs</u>

Large language models come with their own biases. Knowing what these are in regards to this conflict is very difficult to suss out. Indeed, I can be unaware of my own prejudices making it difficult to spot when both the LLMs and my own biases converge.

By using two LLMs I hoped to limit this effect. One example, where I saw a possible bias was with ChatGPT continuing to conflate Hamas with Palestine. So, when an article was highly critical of Hamas, ChatGPT tended to rate it as highly sympathetic to the Palestinians. I would have to often “retrain” ChatGPT with the following prompt:


CH

You

One important thing to keep in mind is that Hamas--the terrorist group--should not be conflated with Palestine or Palestinians in general. Think of it this way: a very pro-Palestine article, written by a Palestinian can be critical of Hamas--the way it has governed, the wanton and unnecessary Oct. 7th attack. The focus should be: does the article acknowledge what Palestinians are currently going through, the history of their suffering, etc. In light of this, how would you adjust your scoring?

ChatGPT

Given this clarification, it's important to reassess the article with a focus on whether it acknowledges the experiences and suffering of the Palestinian people, separate from the actions of Hamas.

Claude showed less biases that I was aware of, and I rarely had to “retrain” Claude. Some of this was derived from Claude’s larger context window (it can “understand” up to 300 pages), but the exchange detailed above took place multiple times within ChatGPTs supposed context window.

Regarding the specific scoring of articles on the scale of -1.0 to 1.0, neither LLM was always consistent. Sometimes, I would have them rescore an article, either in the same thread or in the existing thread, and about 30% of the time, they would offer a different score, albeit one that was within 0.1 of the first score they had offered. This was not too surprising given the complexity of what I was asking the LLMs to do and the fact that my prompt engineering and tuning can subtly influence scoring.

The important takeaway is that the two LLMs were often mostly consistent and when they were not the discrepancy was small.

Finally, I wondered how likely the LLMs were to push back against my own scoring. What if told them that their score was off and then drastically changed the score? Would they go along with me? So, what was not clear was how much my subjective scoring was actually being reined by the models and how much was simply being echoed by the models.

Sentiment Analysis

The difficulty I had working with more advanced NLP tools like BERT limited the validity of this project: they would have provided a far more objective assessment than what I used here. That said, I hope to mitigate this by using two LLMs and having them push back against my scoring to limit my objectivity.

Given more time and greater processing power, I would have explored more tools beyond TextBlob.

Regarding TextBlob, the two tool it gave me did not correlate with what I was looking for: political leaning. They focused more on subjectivity of the writer, something that could manifest on either side of the spectrum. Nonetheless, I decided to include two metrics from TextBlob–'polarity' and 'subjectivity' to see how they interacted with the target variable–'Sympathy Score.'

Difficulty in quantifying bias

Putting a number to something as complex as a political sentiment is understandably a slippery and suspect undertaking, even if there were a panel of experts involved. Here, the number is to a large extent based on my own subjectivity.

One way I tried to limit my own subjectivity was asking the LLMs to push back my own criteria, so that the final criteria used was a mix of their feedback. Granted, how I initially framed the feedback in of itself might limit and define how the LLMs pushed back. My hope was that using two LLMs would limit this effect.

This scale seems to capture a broad spectrum of possible biases in reporting and opinion pieces. However, there are a few considerations that could further refine your approach:

- **Subjectivity in Interpretation:** The scale relies on the reader's interpretation of what constitutes sympathy or bias, which could vary greatly. Establishing clear criteria for what constitutes a -0.8 versus a -0.2, for example, might be necessary.
- **Nuanced or Mixed Viewpoints**: Some articles might present a more balanced view or acknowledge valid points on both sides. It's important to consider how such articles would be scored.
- **Focus on Specific Themes or Narratives**: Some articles might focus on particular themes (human rights, historical grievances, security concerns, etc.) that don't neatly fit into a sympathy scale but still convey a certain bias.
- **Differentiation Between Criticism and Sympathy**: Criticism of actions or policies of Israel or Palestine is not necessarily an indication of sympathy for the other side. It's important to distinguish between critical analysis and outright sympathy or bias.

Your scale is a good starting point, but these considerations could help in applying it more effectively and objectively to a diverse range of articles.

Here are a few thoughts on your proposed scoring system:

- The overall concept makes sense - using a scale from -1 to 1 to capture how sympathetic an article is to Israel vs Palestine.
- The descriptions you provided of what a very pro-Israel (-0.8) or pro-Palestine (+0.8) article would look like are helpful examples.
- I would suggest being a bit careful about using charged words like "oppressor" and "oppressed" in the positive 0.8 example, as those imply a moral judgment, whereas the scoring system aims to capture sympathy levels more objectively.
- For articles close to 0, rather than saying they "downplay" extremes, maybe they could "attempt more balanced perspectives" or "acknowledge suffering on both sides."
- One aspect I'm not totally clear on is how an article would score around 0 specifically. For example, would a purely factual news report with no obvious sympathies score 0? Or would 0 still imply some attempt at a balanced opinion piece?
- Also, for the scale, I'm wondering if articles would realistically score in the extremes like -0.8 or 0.8 very often? Maybe the scale could be adjusted to reflect more likely scores clustering nearer the middle.

Finally, there was the possibility that my own opinion was changing throughout all the reading I did or by what had actually happened that day on the news. I could be subtly influencing the LLMs without being aware that I was doing so. One check I used was to re-score the original five articles I had initially scored when training the two LLMs. I did this about a week after originally scoring them, so I could not recall the number I or the LLMs had given. The results were that in three cases I got the exact same score and the biggest difference was 0.15.

What this all means is that the findings in the EDA section below should not be seen as an exact number. The methodology, my subjectivity, and the lack of exactitude in the LLMs makes variation inevitable. But the variation is mostly minimal and so, in broad strokes, what follows below paints a general picture of media biases relating to the Israel-Palestine that started on Oct. 7th, 2023.

Exploratory Data Analysis (EDA)

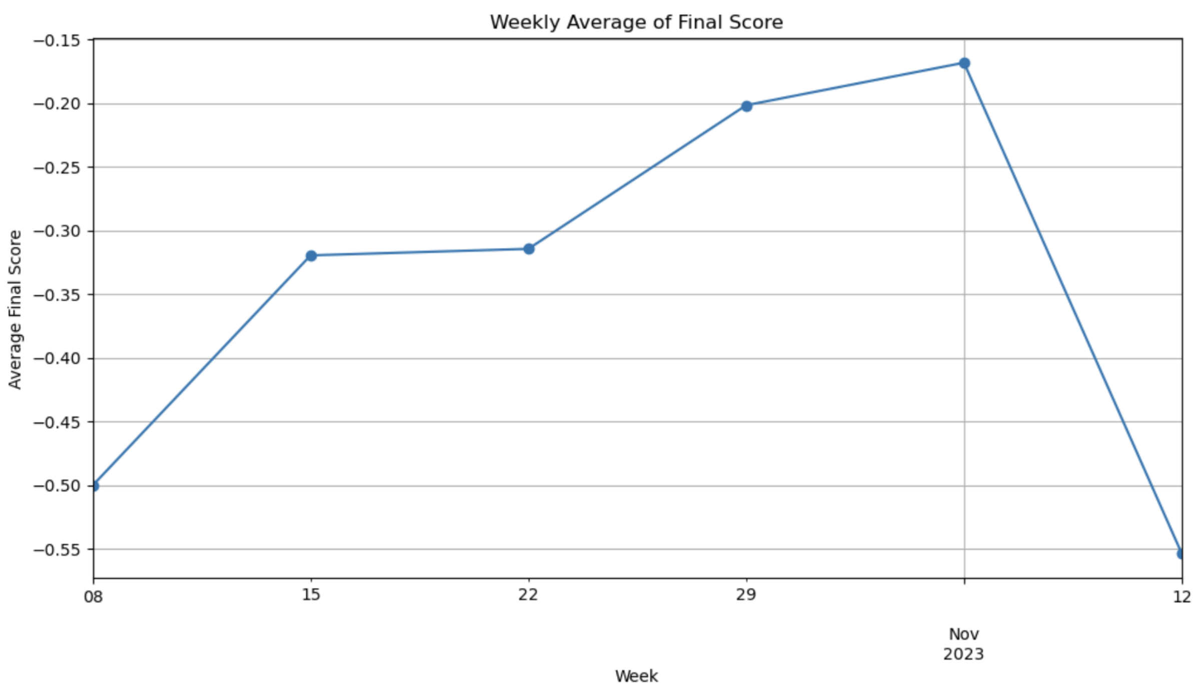
Major Takeaways

- All three media outlets had a pro-Israel bias, ranging from slight (NYTimes.com) to pronounced (WSJ.com)
- NYTimes.com was the most balanced, having a similar mix of Pro-Palestine and Pro-Israel articles
- WSJ.com is the most extreme outlet, publishing the highest % of Op-Eds > than 0.5 or less than -0.5
- All three sources showed marked changes, especially around week 3 and 4, when scores were the most pro-Palestine

I will deal with each of the four major takeaways in its own section below.

<u>Pro-Israel bias in the three sources</u>

Over the course of 5 weeks, averaging out the “Sympathy Score” for the five publications revealed that at no point did the average lean Pro-Palestine. There was also a large range in this sentiment, going from a moderate -.55 (a -1.0 is extremely supportive of Israel) to lightly pro-Israel (slightly less than -.20).



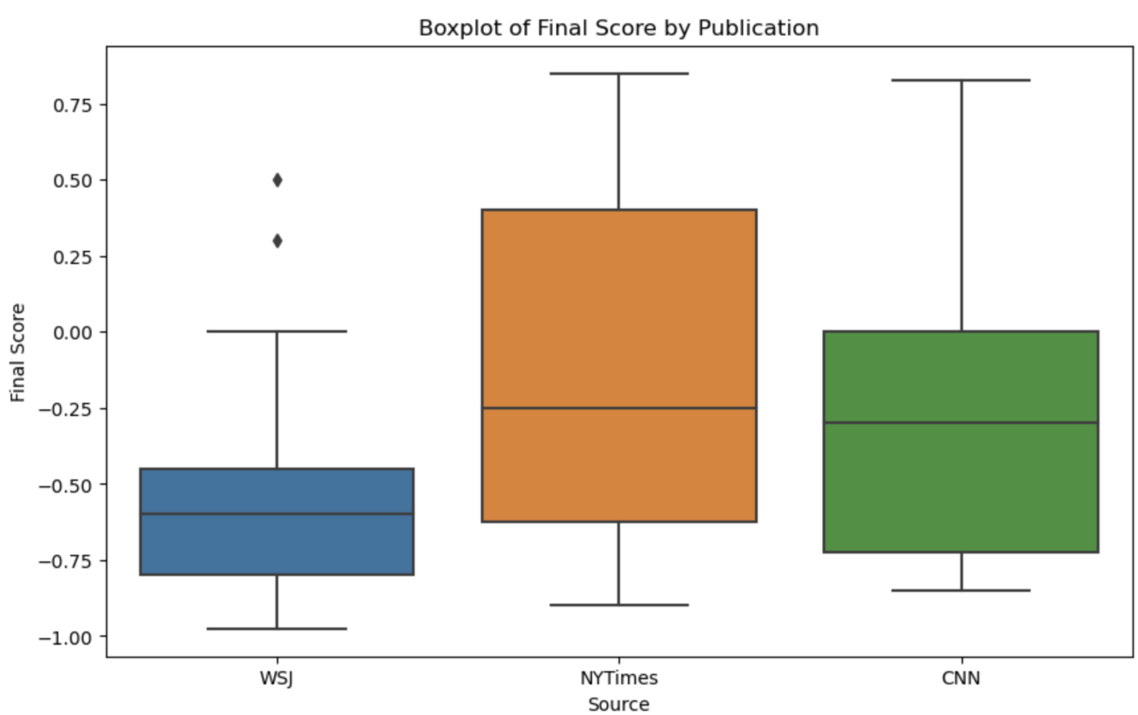
The initial -0.50 “Sympathy Score” across the three publications can likely be attributed to the fact that this was the day that the war broke out in Israel, where a surprise attack on civilians occurred, claiming the lives of over 1,200 and the kidnapping of yet another 240. Given that the “Sympathy Score” was based on how much it acknowledged the suffering of each side, it makes sense that Israel's suffering was what captured the media's attention for that week. And the suffering of Palestinian, which has been taking place over many decades, was not accounted for.

What is not as clear is what happened in the final week (week 5) to account for the strong pro-Israel sentiment yet. After all, the ground invasion of Gaza had begun, as well as the shelling of many civilian installations, including hospitals. The headlines were covered with heartbreaking images of suffering, often women and children (whether Hamas was hiding weapons or secreting its top officials in such installations is an issue that is beyond the intent and scope of this paper, as well as my limited knowledge of the conflict). Therefore, it would seem that pro-Palestine sympathies would be at their highest.

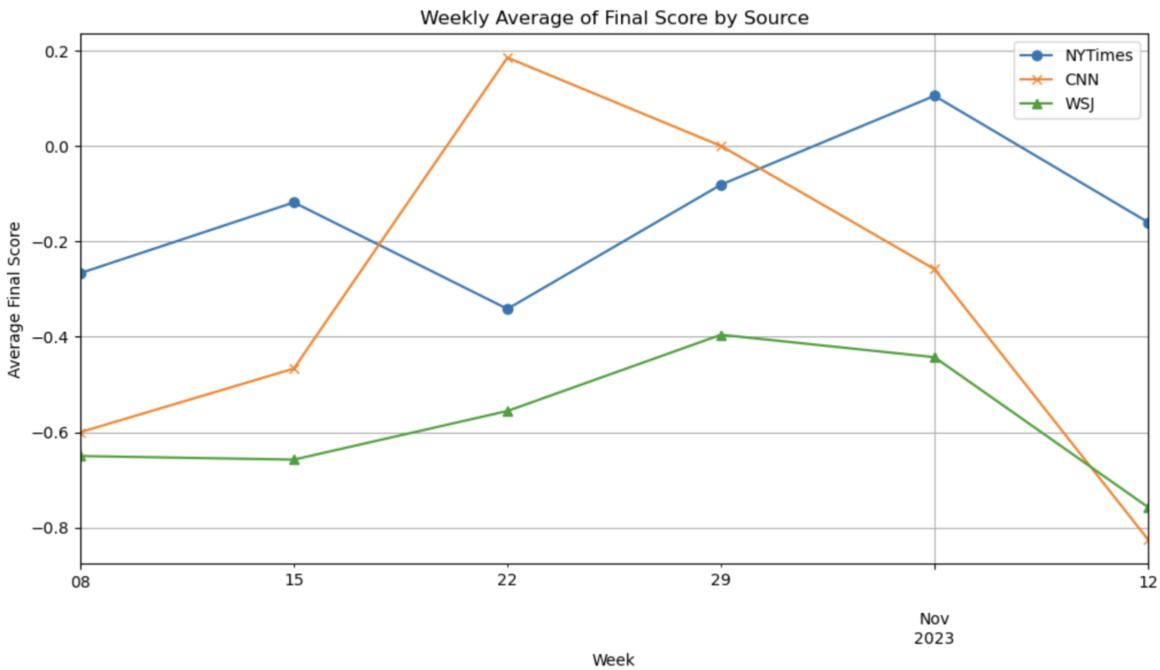
Had I done a sentiment analysis of news articles or looked at what % reported what was happening in Palestine vs. Israel, then they could very well capture what was happening on the ground.

Where Op-Ed pieces for this time period diverged is that much of them related not to the ground coverage but to the growing rise of anti-Semitism on college campuses and in politics. There was also significant criticism of those on the left who supported Hamas (again, whether or not those allegations are true is beyond the scope of this paper). These occurred mostly in the WSJ.com but such articles were also found in the NYTimes.com and CNN.com.

In terms of segmenting between the articles, the boxplot below gives a high-level overview.

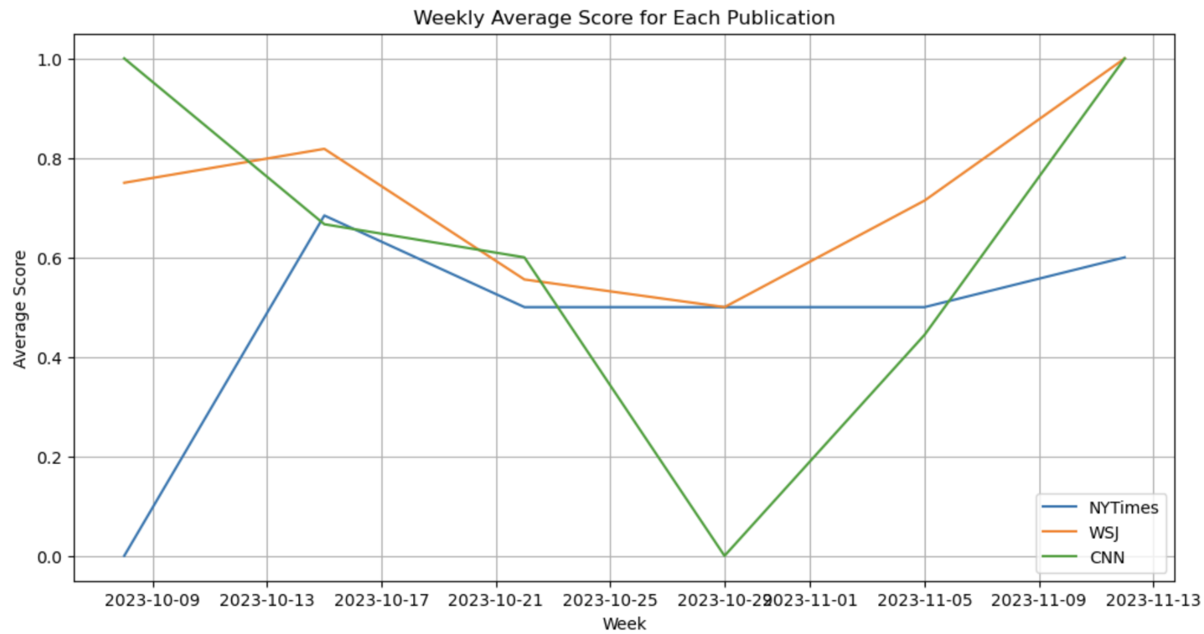


Here we can see that NYTimes covered the largest swath of sentiment, whereas WSJ.com was the narrowest. In terms of average positive score, we can see that a large number of Op-Eds were pro-Palestine in NYTimes, something we don't see for the other two publications. That said, CNN.com did skew pro-Palestine for one week, but that was more on account of a pro-Palestine article skewing the results during a week in which few Op-Eds related to the conflict were published in CNN.



In this time series graph, we can also see how "Sympathy Score" changed for each publication week by week, in a way that was roughly similar. This can likely be attributed to the geopolitical trends: the Oct. 7th attack elicited great sympathy for Israel, the week 3 and 4 trend towards pro-Palestine can likely be attributed to the many Palestinian civilians, including a large number of women and children, who lost their lives in response to Israel artillery, and finally week 5 in which the focus shifted, despite the ongoing off Gaza by Israel, to the spread of anti-Semitism and the condemnation of those in America who support Hamas.

<u>Extreme views in the three media Op-Ed sections</u>



Regarding the Israel- Palestine, in order to account for how extreme the Op-Ed section was I derived an “Strong Views” score, which measured the % of Op-Eds that were either < -0.5 (moderately to strongly pro-Israel) to > 0.5 (moderately to strong pro-Palestine*).

A score of 1.0 in the graph indicates that articles were only extreme. A 0.0 means there were zero extreme articles.

What we see is that CNN had the most variability, coming across as either a highly opinionated Op-Ed section or one that is very neutral. Some of this has to do with the fact that the CNN dataset included only 21 articles. Segmenting those over weeks results in a week with only a few articles, meaning that two extreme articles out of a total of 3 would lead to a score on the graph above of 0.66.

For WSJ.com, we see that they consistently have strong views in their Op-Ed section. Combining this with the information in the boxplot in the section above, we can see that all of these strong views are pro-Israel.

NYTimes.com, by contrast, offers weeks in which strong views represent about half of the Op-Eds published, though there are some weeks where the % of strong views in Op-Eds is relatively limited. What is interesting is that in the week or so after the horrific Oct. 7th attacks the NYTimes had a low “Strong Views” score compared to the other two publications, which were both near the maximum of one. Though in the ensuing weeks, the “Strong Views” scores mostly converge.

*This should in no way be conflated with pro-Hamas. Indeed, some of these articles were anti-Hamas.

<u>"Normalized Diversity" of Scores</u>

The graph above only shows how strong views were, not necessarily which side, if either, they hew to. In order to account for how even-handed an Op-Ed section was I devised another metric: "Normalized Diversity". This indicates the balance between the % of articles that are pro-Palestine and those that are pro-Israel. A really low score indicates that the views are only in one direction. The average per publication is as follows:

```
Normalized Diversity Score by Publication:
Source
CNN        -0.428571
NYTimes    -0.280702
WSJ        -0.840909
```

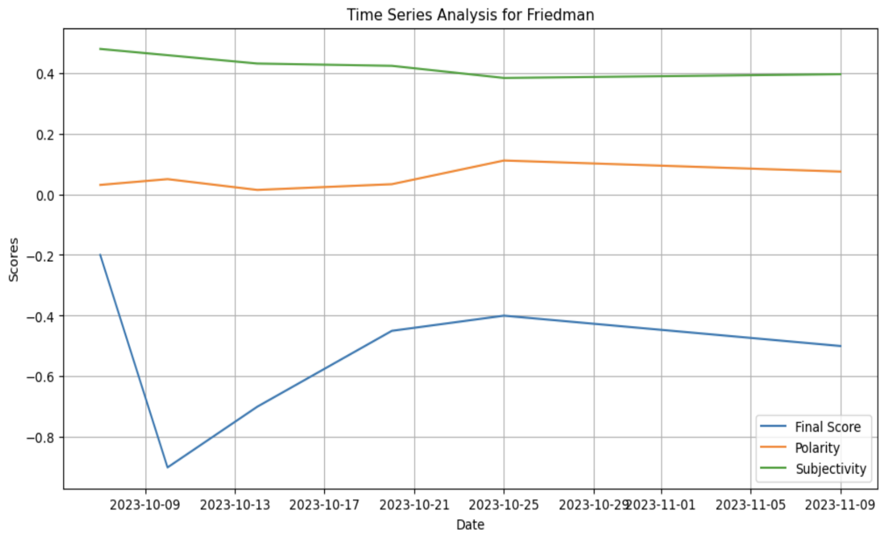
Again, WSJ.com has the least number of articles that are sympathetic to the other side. NYTimes has the most that are sympathetic to both sides. CNN falls somewhere in the middle.

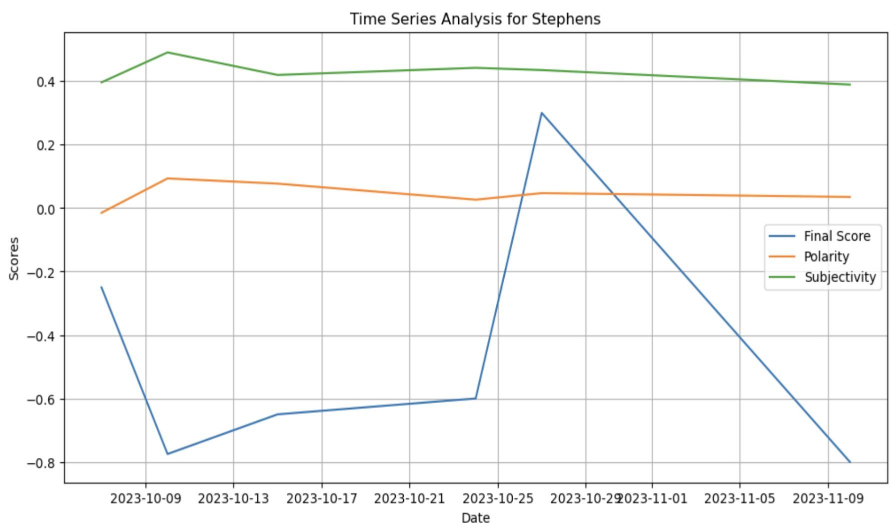
The following graph shows how this plays out over the timeline covered in this project.

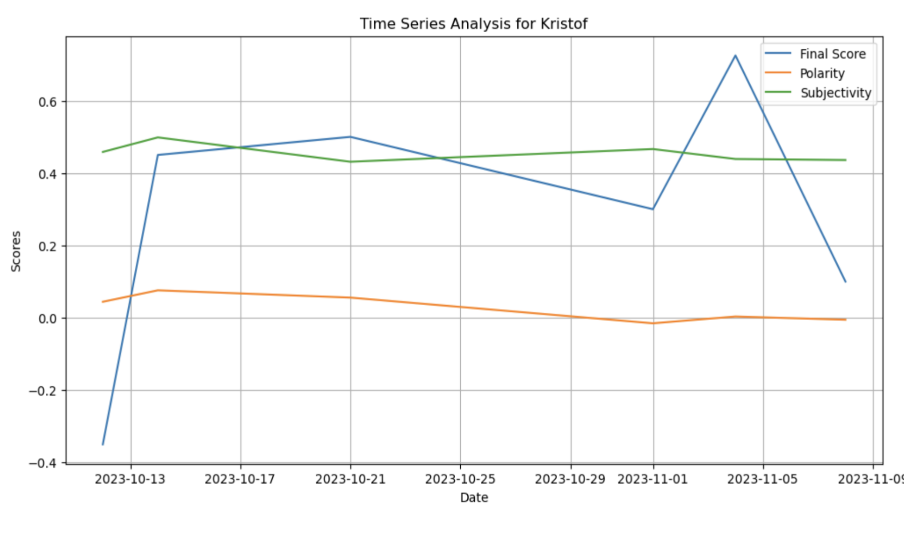
<u>Additional observations</u>

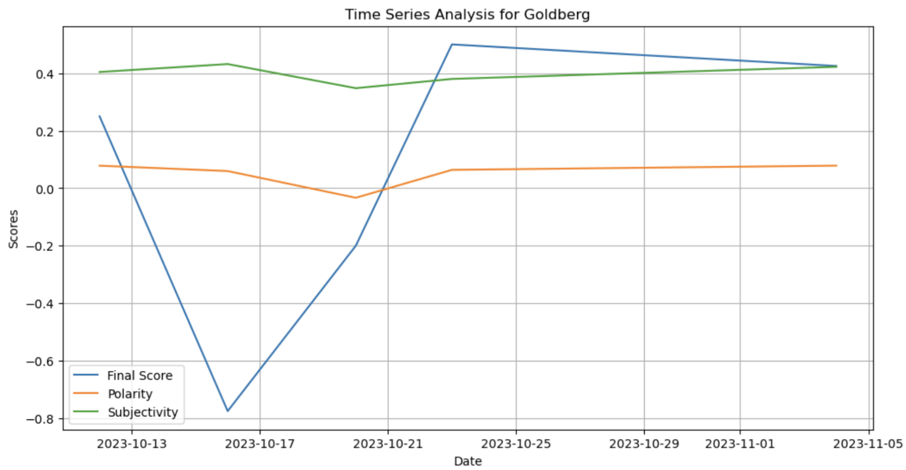
One thing I also explored was whether the "Sympathy Score" for the same author changed over time. I was trying to see whether the rapidly shifting geopolitical realities might influence a writer's sentiment.

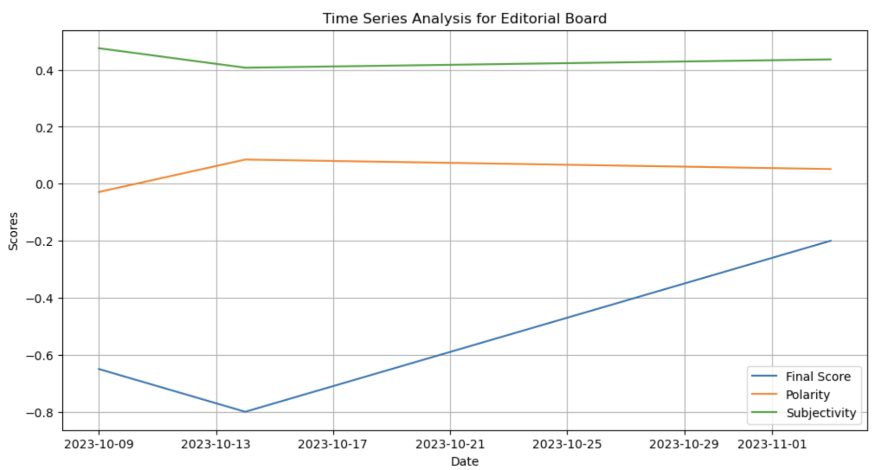
Given the relatively small dataset–124 articles–there were only a few authors who published 3 or more articles. Nonetheless, there were some interesting trends as seen below.


Time Series Analysis for Friedman
Final Score
Polarity
Subjectivity
Scores
Date
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
2023-10-09
2023-10-13
2023-10-17
2023-10-21
2023-10-25
2023-10-29
2023-11-01
2023-11-05
2023-11-09


Time Series Analysis for Stephens
Final Score
Polarity
Subjectivity
Scores
Date
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
2023-10-09
2023-10-13
2023-10-17
2023-10-21
2023-10-25
2023-10-29
2023-11-01
2023-11-05
2023-11-09


Time Series Analysis for Kristof
Final Score
Polarity
Subjectivity
Scores
Date
0.6
0.4
0.2
0.0
−0.2
−0.4
2023-10-13
2023-10-17
2023-10-21
2023-10-25
2023-10-29
2023-11-01
2023-11-05
2023-11-09





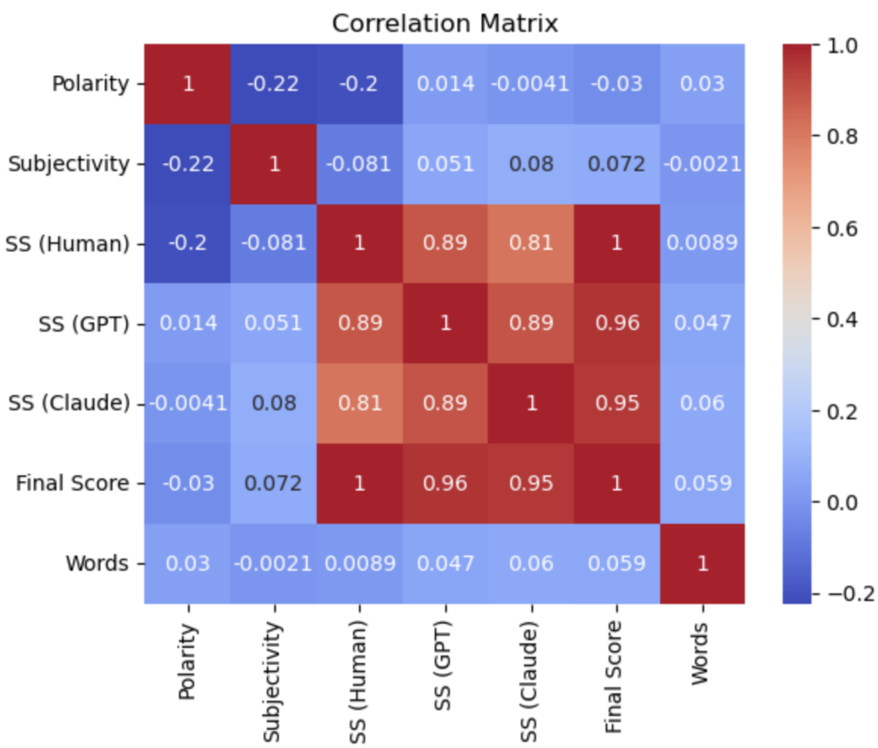
The following graphs were for NYTimes writers. We can see that all the writers had a surprising variability in their “Sympathy Score”, or “Final Score.” My expectations were that these writers would likely be more consistent in their views and that a diversity of “Sympathy Scores” would come from an Op-Ed section having different writers, not necessarily from the same writer.

This variability was in fact strongly pronounced. Anna Goldberg (second graph from the bottom) had a Sentiment Score that differed by nearly 1.3 out of a total of 2. To put this in another context, though still a political one, it as though your friend was one week talking fervently about their support for, say, the Republican party and then two weeks later claiming that they are a moderate Democrat.

Of course, the analogy is not perfect given how quickly events in this conflict have shifted. Notably, there was only one deadly attack on Israel (there were a few more thwarted attacks), whereas the bombing in Gaza has been steady and intense throughout the weeks. My theory is that, to some extent at least, this dynamic has accounted for shifting sympathies, as writers become more sympathetic to the suffering in Gaza. That said, as noted in the section above, that sentiment shifted as many Op-Ed pieces began to focus on anti-Semitism and support of Hamas.

Polarization and Subjectivity

Interestingly, there was essentially no connection between these two scores (generated by TextBlob) and “Sympathy Score” (or Final Score).



Both of these account for aspects of the text that don’t speak to political leaning or sympathies. Polarity essentially indicates whether a text has a positive, negative, or neutral sentiment, and subjectivity to what extent a piece is neutral. Given these were all Op-Ed pieces, subjectivity is going to be high and there is likely to be sentiment, either negative or positive. That neither side of the political spectrum, at least as defined by these three Op-Ed sections, is differentiated by their polarity or subjectivity is perhaps not surprising.

This outcome, I believe, speaks to the power of coming up with something like a Sympathy Score: it allows for another way to do political sentiment analysis for when the NLPs tools at hand are unable to do so.

Modeling

Baseline model
To train the model I started with Ensemble methods. Though the dataset was relatively small, my thinking was that Random Forest is a good way to establish a baseline, after which I can try other techniques. I was also concerned about overfitting and thought that Random Forest might be able to control, at least somewhat, for this.

The following numbers were obtained:

Model MSE: 0.34134798999999993
Model RMSE: 0.5842499379546393
Model MAE: 0.46062
Model R-squared: -0.022134622525122882

Given that the target variable –'Final Score'-- is based on a range from -1 to 1, a MSE shows that the model is performing inaccurately. This is both concerning, since it is likely making some strongly inaccurate predictions and not concerning (or at least too concerning), since this is the very first model I tried.

Also of note is that the R^2 is performing worse than a horizontal line, albeit very slightly, with a score that is negative.

Sentiment analysis tools

To better capture the textual nuances, I incorporated sentiment analysis tools, starting with Word2Vec, which yielded the following numbers:

Model MSE: 0.367830455
Model RMSE: 0.6064902760968225
Model MAE: 0.48812
Model R-squared: -0.10143388650001794

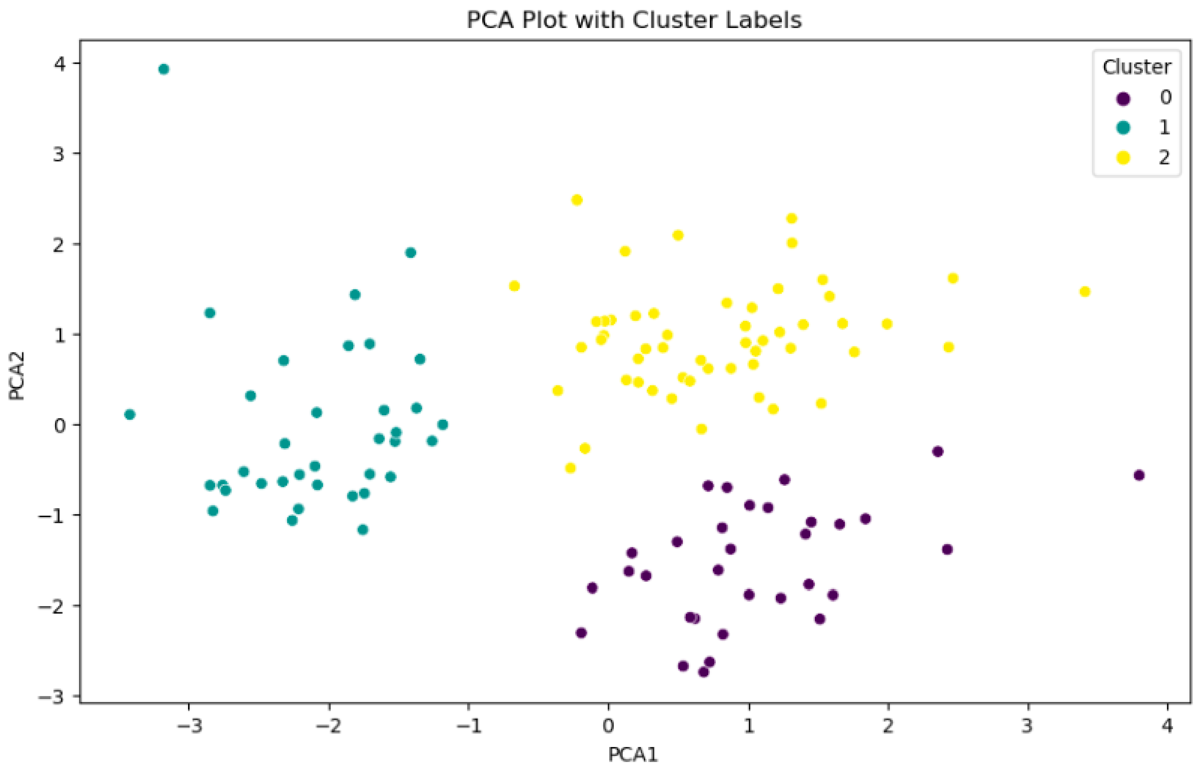
The performance fared even more poorly. Again, the small size of the dataset could likely account for this.

I thought of incorporating other approaches with Word2Vec beyond Random Forest, such as Linear Regression, SVMs, and then hyperparameter tuning. Yet, the performance was so substandard that I opted to change the approach.

Unsupervised Learning

I decided to use K-means clustering with an n of 3, given that there were three different publications. My hypothesis was that there might be differences between the three sources–NYTimes.com, CNN.com, and WSJ.com–that might manifest to some extent using K-means clustering.

Additionally, I incorporated PCA in order to reduce the dimensionality of the space and moderate the effect of variance, which plagued the supervised models employed above. (Additionally, I normalized features to account for word length, since the number of words per article tended to differ between the three publications.)



Here we can see that there are three relatively distinct clusters. To see how these three clusters skewed in terms of the publication, I ran a cross tabulation.

```
Cluster   0   1   2
Source
CNN       4   5  12
NYTimes  10  27  21
WSJ      18   4  20
```

Cluster	0	1	2
Source			
CNN	0.190476	0.238095	0.571429
NYTimes	0.172414	0.465517	0.362069
WSJ	0.428571	0.095238	0.476190

CNN showed the strongest affiliation with cluster 2, meaning that these articles had an almost 60% chance of being in Cluster 2. Whereas for Cluster 0, there was a greater than 50% chance that the article came from WSJ.com.

Based on my own exposure reading many of these articles during the training periods of the Generative AI models, one hypothesis is that Cluster 1 captures a pro-Palestine slant. Only 2 articles in WSJ were not pro-Israel, reflecting the 9.5% number for Cluster 1. Cluster 2 could represent moderate pro-Israel sentiment, which was common on all three publications. In fact, throughout the entire five weeks of observation, there were a total of two weeks across all three publications in which the average sentiment was pro-Israel (so 2 weeks out of a total of 15). This is consistent with the fact that cluster 2 accounts for the majority of the data points across all three clusters.

Finally, Cluster 1 could be strong pro-Israel sentiment, given that WSJ accounted for a majority of these pieces in the analysis.

Wrap-up

Given time constraints, I was not able to pursue modeling beyond this. My plan was to feed text for articles for week 6 to see which of the three sources the article was most likely from. Then, I would compare this with the performance of the GenAIs' predictions (see section below) for those very same week 6 articles.

Additional Application of GenAI Models

In order to predict which of the three sources a week 6 article came from, I also employed Generative AI. Below are the following steps I used:

1. Feed an article into one of the existing threads in which had been trained and calibrated on rating the article.
2. Ask it to identify which of the three sources an article is most likely to come from.
3. Provide one (and only one) round of feedback/calibration after the first misidentified article.

The goal here was to have a streamlined process that would be a relatively low level of effort to incorporate if subsequent weeks are included. To “test run” this process with the GenAI, I used eight articles from week 6 (the study was conducted from week 1-5, so week 6 was not included in the original dataset nor was available at the time of the creation of the data set.

Here are the results:

LLM	**Success Rate**
ChatGPT	6 of 8 articles/75%
Claude	6 of 8 articles/75%

Of note, of the two that the LLMs missed, both missed the very first article, after which I provide feedback and calibration. Like the “Sympathy Score” itself, this suggests these models are trainable. Additionally, the two LLMs both miscategorized the same article post-calibration. This was a NYTimes Op-Ed, which they both categorized as coming from WSJ.com. The author, Bret Stephens, is a former WSJ.com writer and editor whom the NYTimes brought on to provide a diversity of viewpoints.

I do not expect outliers like this to be picked up in the model, unless there are some parameters at play in the NYTimes.com Op-Eds (e.g., word length, use of anecdote(s), etc.) that the model is not picking up on.

Wrap-up

This project provided insights around the different editorial biases of the three Op-Ed sections, as these biases pertain to the Israel-Palestine conflict. Here are the major learnings:

- Both the NYTimes.com and CNN.com Op-Ed sections were much more likely to have Op-Eds that had both pro-Palestine and pro-Israel articles, as defined by the “Smypathy Score” (also referred to as Final Score).
    - This was captured using a Diversity Score in which the number of pro-Israel articles was subtracted from pro-Palestine articles.
        - In terms of a specific week, WSJ.com had the highest score of -10 and CNN the lowest with 0
    - During the first two weeks of the conflict, the scores were the least diverse
        - Likely because Israel had been attacked and had yet to respond significantly

    - These are both left leaning publications, yet showed a richer diversity of viewpoints in relation to the Israel-Palestine conflict.
        - The NYTimes.com was judged as the farthest left, though regarding this specific conflict, it actually trended slightly right.*

- In terms of weekly changes the Op-Eds from all three sources skewed pro-Israel, whereas in weeks 3-4 they trended pro-Palestine, before trending pro-Israel in the final week.
    - The extent of the pro-Israel score across all three sources was stronger than that of the pro-Palestine score
        - NYTimes.com and CNN.com registered a pro-Palestine “Sympathy Score” but only for a total of 2 weeks and the effect was very slight.
    - Geopolitical factors can likely account for some of the changes, including an Israel military response that occurred during the latter weeks documented here
        - The final trend towards Israel can likely be attributed to an increased focus on anti-Semitism domestically and abroad.

*The US government’s and by extension the US media’s support of Israel is complicated and has a long history. Therefore, to equate left with not being supportive of Israel and right as supportive is overly simplistic and does not accurately capture the reality.

<u>Use of GenAI</u>

Article evaluation
GenAI was a fundamental aspect of this project. Embarking on this project, it was not immediately clear how well the GenAI models would do. One scenario that could have played out was that the LLMs were highly inaccurate and not amenable to calibration. However, it only took five scorings of Op-Eds to get both models to be within 0.1 of my own scoring. Once this threshold had been established, the project flowed relatively smoothly, with the LLMs diverging by more than 0.25 in 21/124 cases or about 15% percent of the time.

There were some issues, though. ChatGPT due to its smaller context window presumably, needed more reminders, which necessitated the reposting of the criteria. Sometimes this happened after only a few articles, so clearly context window size was not an issue. One observation was that this seemed to happen the “deeper” the ChatGPT thread. This “context drift”, as I’m dubbing here, is something that OpenAI announced they will be improving in their soon-to-be-released model–ChatGPT 4 Turbo.

In this regard, Claude was the superior LLM. I only had to change threads once (vs three times with ChatGPT), and I do not recall one instance of “context drift.”

In terms of accuracy, where accuracy is defined as how close the LLMs "Sympathy Score" was to my score, the two models were nearly identical with Claude doing better in 10 instances and ChatGPT doing better in 9 instances. There were two "ties."

Article Identification

The sample size here was very small, so it is difficult to draw any definite conclusions. Not including the first "training" article, the models both correctly identified the Op-Ed source, 6 out of 7 times, the one miss being attributable to the fact that the NYTimes writer was a former editor and columnist at the WSJ.com. These are promising numbers and speak to the potential use cases of LLMs to quickly identify, based on language and nuances, the origin of a text source.

Given that I was not able to leverage sophisticated NLPs, such as BERT, using LLMs is relatively low effort and provides, from the small dataset, potentially highly accurate predictions. That said, I did use Word2Vec and that did not yield a level of prediction comparable to what the LLMs achieved (though the use case here was a little different because the LLMs were picking one of three sources versus using sophisticated predictive analytics.)

<u>Future applications of GenAI</u>

I do not expect Generative AI to replace NLP tools when doing sentiment analysis. Though without seeing how a sophisticated model like BERT would have performed, it is hard to make a definite conclusion. However, if someone has limited computational power or does not possess the requisite coding skills, then using ChatGPT, Claude, or another LLM to do sentiment analysis is viable. Not only that, but it is possible to highly customize the exact sentiment you are trying to detect.

That is not to say there are no potential issues with a human in the loop. Firstly, the prompt input supplied by the human must be clear and the human must be able to engage in training and calibration, something that doesn't take place in one prompt but that often plays out over several prompts. Furthermore, the human must be open to refining their own instructions based on the feedback from an LLM(s).

In regards to LLMs, they too have their issues, from hallucination to limited context window to, what I've dubbed, "context drift." Using at least two LLMs can mitigate these effects and provide a more valid and robust way of doing sentiment analysis.