Quantifying Protest Framing: Large Language Models for Sentiment and Paradigm Scales

Sophia Tannir

April 2023

1 Introduction

In recent years, large language models (LLMs) have emerged as great tools for various natural language processing tasks. Building upon the work of Wu et al. (2023), who used LLMs to estimate the latent positions of politicians, my capstone proposes a novel method for scaling news articles based on their framing of protests and the causes behind them. While Wu et al. (2023) leveraged the information already embedded in pre-trained LLMs, this method involves feeding the LLM a particular comparison, the full text that should be compared, and then requiring the model to make a selection.

By using the analytical capabilities of LLMs through pairwise comparisons and the Bradley-Terry model, I hoped to create continuous, unidimensional scales that capture the latent dimensions of how news articles portray protests and cover (or do not cover) underlying causes. Moving forward, these scores can then be used as inputs (or outputs) for models such as regression when analyzing protest coverage and potential underlying biases in media.

2 Literature Review

Media coverage plays a crucial role in shaping public perception of protests and social movements in general. Analyzing this coverage is critical, but traditional approaches often rely on manual tagging and content analysis, which is time-consuming and resource-intensive. These limitations have led researchers to explore new methods, particularly large language models, to automate the process of classifying texts enable larger-scale analysis.

2.1 Framing of Protests in Media Coverage

Many studies have investigated how media coverage portrays protests and the protesters themselves. For this project, I had a particular interest in the coverage of protests that "challenge the status quo", as well as the factors that influence the media portrayals of these types of protests.

Brown and Harlow (2019) find that the choice of protest topic and region can shape media coverage, with protests centered on racial issues more likely to be delegitimized, while Weaver and Scacco (2013) find that even right-wing movements like the Tea Party can be subject to marginalizing frames, particularly from ideologically opposed news sources. Gottleib (2015) takes a longitudinal approach, examining how The New York Times' framing of the Occupy Wall Street movement evolved over time, shifting from a focus on economic concerns to the escalating conflict between protesters and city officials as the movement reached its peak. I used this differentiation between a causal and demand-based framing of a protest as opposed to a tactics-based frame to formulate my second research question, drawing on the protest paradigm introduced by McLeod and Hertog (1992).

2.2 Traditional and Machine Learning Approaches to Text Classification

Machine learning techniques as classifiers for natural language processing (NLP) tasks have emerged as the preferred method for large-scale data problems. Miric, Jia, and Huang (2022) demonstrate the application of supervised ML methods to classify patent abstracts based on their relationship to AI technologies, providing an overview of the considerations and trade-offs associated with different ML approaches. Supervised ML inherently comes with a labeling problem, requiring labels to be generated before the models can be actually applied.

Another approach to training classifiers for natural language processing tasks is proposed by Hancock et al. (2018), who introduce BabbleLabble, a framework that uses natural language explanations provided by annotators to generate programmatic labeling functions. These labeling functions can then be used to generate "noisy labels" for large amounts of unlabeled data, enabling faster training of classifiers with comparable performance to traditional labeling methods. However, these labels are still considered "noisy" and will require further analysis. This implies a trade-off between the time-intensive approach of hand labeling large amounts of data or sacrificing data quality while still requiring human annotations.

2.3 Usage of LLMs for NLP Tasks

Wu et al. (2023) demonstrate the use of LLMs in addressing complex measurement problems, using them to estimate the latent positions of U.S. senators. They leverage the already embedded knowledge in LLMs to make pairwise comparisons between senators and then scale the results using the Bradley-Terry model. While they compare politicians and my capstone focuses instead on news articles, their findings suggest that LLMs can be used to generate this dataset of comparison selections. By building on their novel approach and applying it to the problem of scaling protest framing in news articles, I've changed the method by feeding the LLM the actual text to be compared instead of using the already-existing knowledge in the LLM. I then finish by using the same scaling method, the Bradley-Terry model.

3 Methods

NLP tasks often rely on either unsupervised methods like topic modeling or supervised approaches that require extensive human annotation and hand-coding. While unsupervised methods can uncover certain patterns in text data, they often struggle to actually capture the specific dimensions of interest. The process of manually coding datasets is time-consuming, expensive, and prone to human biases. Cognitive constraints on human coders hinder the analysis of subtler or more exhaustive categorizations of text that might be of scholarly interest, as the mental effort required to analyze and categorize large volumes of text can lead to burnout and increasingly poor results. Recent advancements in LLMs offer a new solution by automating the process of identifying and categorizing subtle aspects of text that are of particular interest to specific research questions. I leverage that solution here.

3.1 Data Collection

I collected a dataset of 140 news articles from the New York Times using the NYT API. The articles were retrieved using the query term "protests" and were published within the past two years from the current date. The abstract, lead paragraph, and headline of each article were extracted and combined into a single text for analysis. This ensures general consistency across articles and controls for potential differences in article length.

3.2 Pairwise Comparisons using LLMs

To motivate the method, consider the following two examples of New York Times articles about protests.
EXAMPLE 1: TACTICS-BASED COVERAGE

- Headline: Conservative Influencer Is Charged in Jan. 6 Attack
- Abstract: Isabella DeLuca, 24, helped to steal a table that rioters used to assault law enforcement, according to a criminal complaint.
- Lead Paragraph: A conservative social media influencer has been arrested on misdemeanor charges related to her involvement in the Jan. 6, 2021, riot at the U.S. Capitol, including an accusation that she helped to steal a table that the F.B.I. says was used to assault officers, according to court documents.

The first example illustrates an article that is about the methods used by protesters, one protester in particular here. By highlighting the details of her actions, the article draws attention to the violent and destructive tactics employed by some of the protesters during the event. The article is also relatively negative in tone, focusing on the criminal charges brought against DeLuca and the violence that occurred during the protest. It does not delve into the underlying motivations or grievances of the protesters but instead concentrates on the unlawful and aggressive nature of their actions. The language used, such as "riot," "steal," and "assault," reinforces the negative sentiment surrounding the event and the protesters involved.

EXAMPLE 2: DEMANDS-BASED COVERAGE

- Headline: Artists and Speakers Withdraw From SXSW Over U.S. Military's Support of Israel
- Abstract: About 80 musical acts, conference speakers and sponsors have pulled out of the festival in Austin to protest sponsorship by the U.S. military, which has lent support to Israel in the war in Gaza.
- Lead Paragraph: Dozens of musicians and panelists have withdrawn from the South by Southwest Conference and Festivals in Austin, Texas, to protest sponsorship by the U.S. Army and defense companies as pressure grows against the U.S. military's support of Israel's war against Hamas in Gaza.

The second example is about the protesters' demands, describing the causes of their discontent and explaining why the protest is taking place. The article explains that approximately 80 musical acts, conference speakers, and sponsors have withdrawn from the festival to protest the sponsorship by the U.S. Army and defense companies. This mass withdrawal is a clear indication of the protesters' strong opposition to the U.S. military's involvement in the Israel-Hamas conflict and their desire to bring attention to that involvement.

While this is not a "protest" in the traditional sense (ie. a street protest), the difference in the focus on demands is clear. The article is more neutral in tone than the first example. While it stops short of explicitly endorsing the protest, it nevertheless depicts the protest in a more positive light.

The preceding examples illuminate the two dimensions of substantive interest: the topical focus of the article and the sentiment of the coverage. These dimensions are of common interest in the study of political communication and media frames. However, as discussed above, measuring these dimensions at scale typically requires prohibitively expensive human coders to be trained on a complex codebook and then spend hours reading the articles in order to annotate the data. Furthermore, human fatigue can set in, especially in settings where each unit of interest is a substantial amount of text. Even with trained expert coders, there is a concern that burn out will systematically reduce data quality.

I instead investigate the performance of an LLM (GPT-4) which is used to select the article in which the dimension of interest is more evident. For example, I ask GPT-4 to indicate which article is more positive, or to indicate which article emphasizes the protesters' demands more. Specifically, I use the GPT-4 model to conduct pairwise comparisons between the news articles. For each pair of articles, the LLM was prompted with the following questions:

1. Sentiment: "Out of the two examples, which article is framing the protesters in a more positive light?"
2. Paradigm: "Out of the two examples, which article is focusing on the protesters' demands and the causes of the protest, as opposed to the protesters' tactics?"

The LLM's responses were extracted and recorded as the "winner" (the article chosen by the LLM) and "loser" for each pairwise comparison. Since I tested all pairwise comparisons exhaustively, we should expect a roughly 50-50 split between winners and losers, depending on the dimension. Ties were allowed when the LLM could not determine a clear winner (tagged as 'None' by GPT), and these rows were excluded from the final dataset, which removes less than 1% of the total comparisons.

Figure 1 plots the probabilities that each article was chosen as more focused on the protesters' demands (left facet) or more positive (right facet) across all results.

3.3 Bradley-Terry Model

While descriptively interesting, we can do more to learn about the latent dimensions of interest. Specifically, I implement a Bradley-Terry model (Bradley and Terry 1952) to estimate the latent positions of the news articles based on the pairwise comparisons. The model assumes that the odds of article i being preferred over article j are proportional to the ratio of their latent "ability" parameters, $\exp(\lambda_i - \lambda_j)$. Originally designed to characterize individual performance (i.e., athletes, students, teams), the Bradley-Terry approach is readily amenable to any setting in which many pairwise evaluations are observed for a single unit.[1] In our

[1]The necessary number of pairwise comparisons require for stability is unclear. In this setting, with only 140 articles in total, the model can estimate the exhaustive set of pairwise comparisons at minimal computational and financial cost. However, implementations with larger datasets might require an increasingly sparse set of comparisons, the impacts of which on the Bradley-Terry model's stability are uncertain. I return to this limitation in the Discussion and encourage future research to characterize this limitation more systematically.

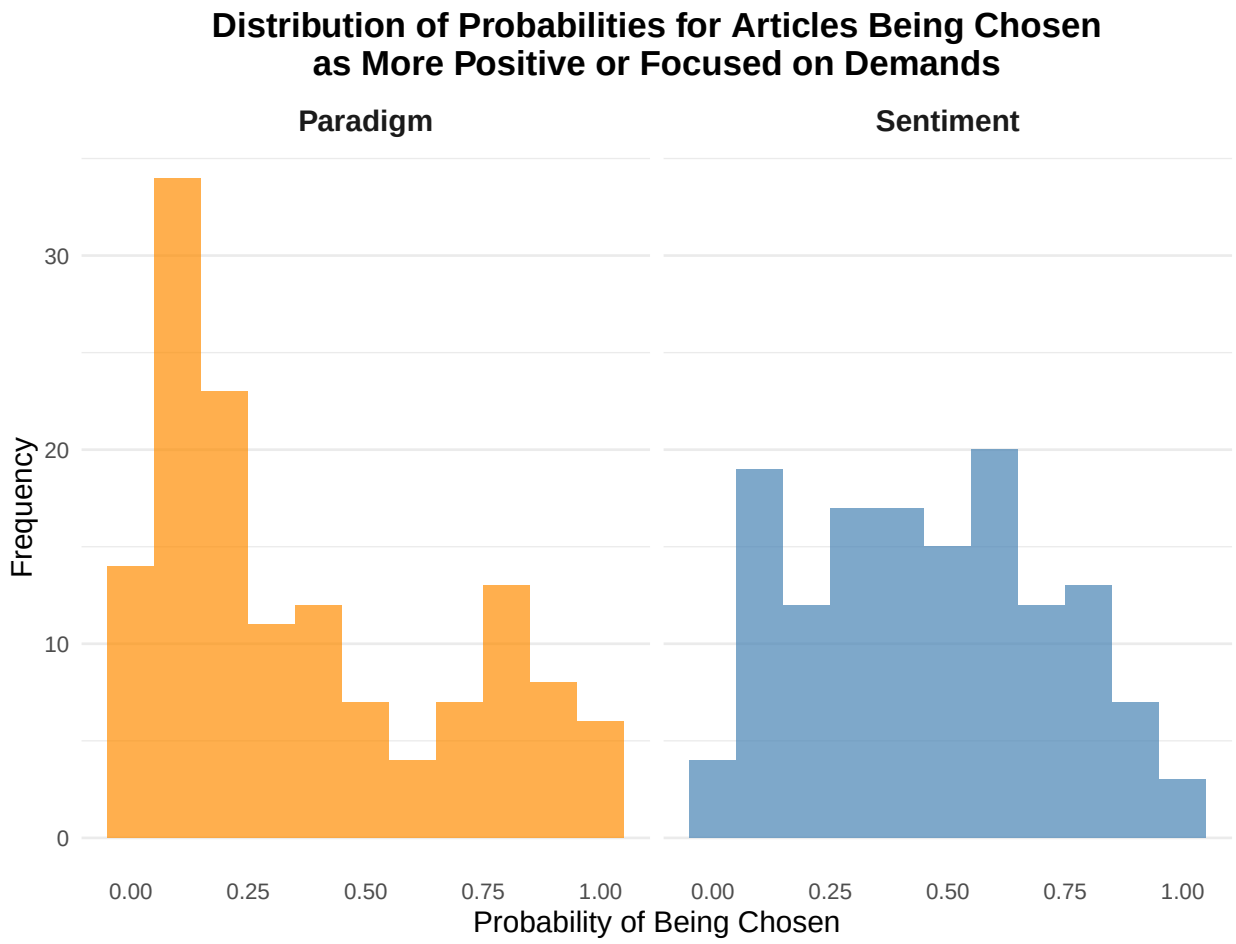

Figure 1: 140 articles' probabilities of being selected as the winner, either for more positive sentiment or more cause-based coverage

setting, estimated λ parameters (i.e., the "ability" parameters) represent the latent positions of the articles on the sentiment and paradigm scales (Wu et al. 2023).

The resulting scores are similar to other latent estimation methods in the sense that the scales are not necessarily substantively meaningful in isolation. However, unlike other latent measures used to estimate things like ideology or extremism, the sign of the Language Model Pairwise comparison (LaMP) scores is interpretable in the context of the pairwise prompts. In our setting, larger values reflect articles which are "more" along both dimensions of interest: positive and about protester demands.

4 Results

I start by describing the estimated dimensions of interest both qualitatively and quantitatively. Qualitatively, I provide evidence that the scales cohere with face validity tests, demonstrating that the articles which score on the extremes of our latent measures indeed reflect the characteristics I would expect. Quantitatively, I demonstrate that both measures are roughly normally distributed, although show that there is more skew in the Demand/Tactics scale than in the net sentiment scale.

4.1 Sentiment Scale

The most positive and negative articles in our data are reproduced below. As illustrated, these examples exhibit strong face validity. The most positive article is about the funeral of Cecilia Gentili, a transgender activist and actress, held at St. Patrick's Cathedral. The article intentionally highlights the mood of the protest, calling it "jubilant" as well as "exuberant". This article emphasizes the packed pews and atmosphere, and focuses on the joy and triumph in honoring Gentili's legacy in regards to the progress made in terms of LGBTQ+ acceptance within the Catholic Church.

Meanwhile, the most negative article discusses the "surge in hostility toward migrants in Ireland, fueled by a housing crisis and far-right influencers." The negative tone is apparent in the language used, such as "hostility," "anger," and "divided," which emphasizes the tense and confrontational atmosphere surrounding the protest itself and the broader issue of anti-immigrant sentiment in Ireland.

Another great example of a negatively-toned article was identified as the second-most negative article. This article focuses on the "violent riots" in Papua New Guinea and the subsequent declaration of a state of emergency by the prime minister. The abstract highlights the unclear extent of damage and casualties resulting from the unrest, immediately setting a tone of chaos and uncertainty. Language includes "shell-shocked", "violent protests," "unrest," "damage," and "casualties." It portrays a situation of chaos, violence, and instability. It also focuses on the impact for authorities as they struggle to restore order in the face of significant unrest.

I plot the distribution of the latent sentiment scores in Figure 2, highlighting evidence of a slight skew in the data. The majority of articles appear to be neutral or negative in their coverage of the protest, with a few exceptions where the coverage is strongly positive.

4.2 Paradigm Scale

Turning to the paradigm scale, I also find strong face validity in the scores. This is based on a reading of the articles that were identified as being most about the protesters' demands. The article that focuses most on the protesters' demands and the cause of the protest is titled "Shaken by Grisly Killings of Women, Activists in Africa Demand Change." This article very clearly defines the cause of the protest as the high rate of gender-related killings of women in Africa. The article goes on to describe the accusations that officials are ignoring the issue and blaming the victims. This article focuses on the underlying causes of the protests and the specific changes the protesters are demanding from their governments, asking for action specifically regarding femicide.

Conversely, the article that least emphasizes the protesters' demands, and instead focuses more on protester tactics, is titled "Does the Peace Sign Stand a Chance?" The abstract suggests that a once-powerful protest symbol, the peace sign, has lost its impact on younger generations, comparing it to a smiley face. The article goes on to discuss changes in symbolism for protesters and the different usages of protest symbols. While this article is not about any specific protest, its scoring highlights the ability of my model to

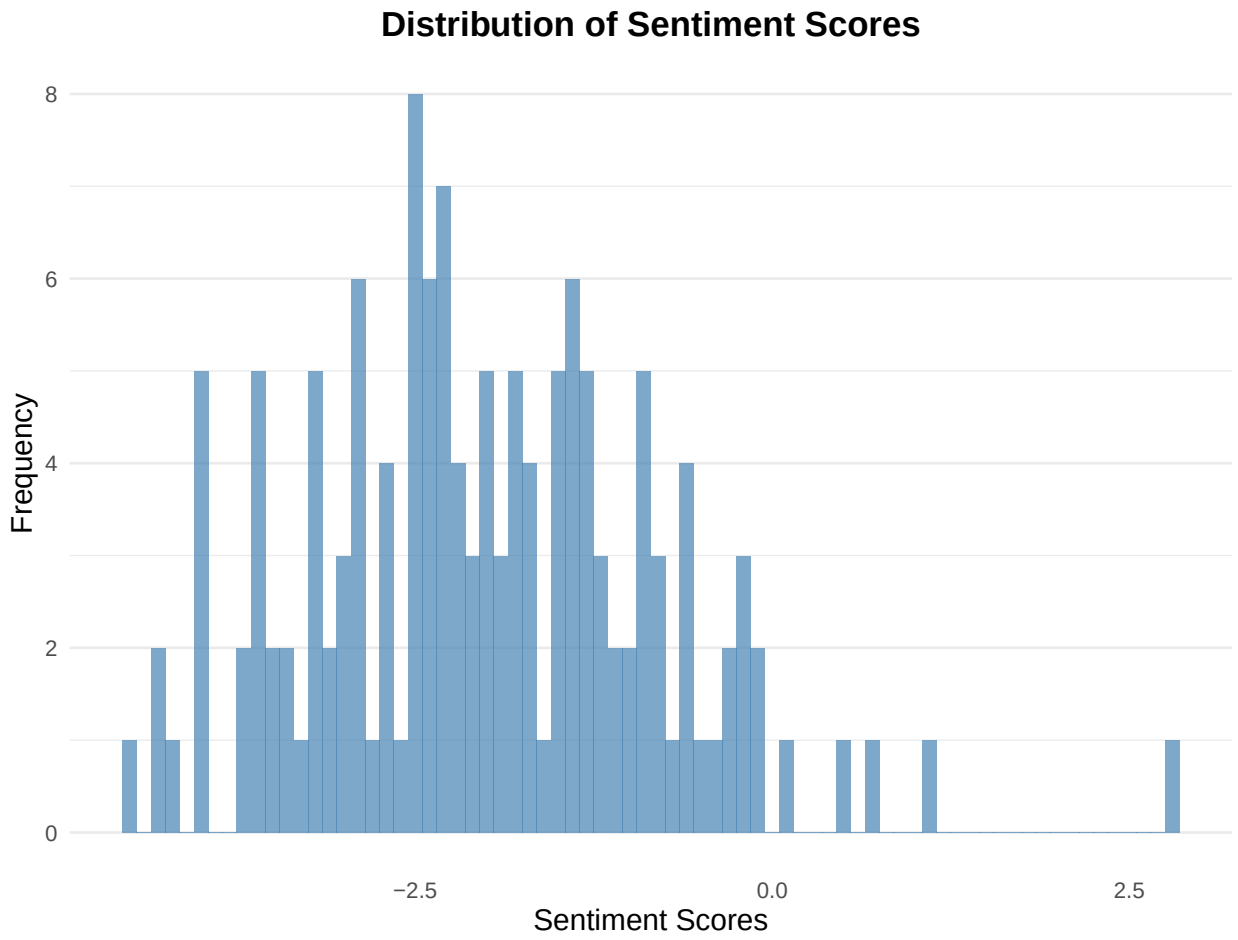

Figure 2: Distribution of sentiment scores derived from the Bradley-Terry model

identify a focus on protest tactics and a lack of underlying cause. As above, I visualize the raw distribution of the paradigm scores in Figure 3. Here, I do not find much skew, showing a relatively even distribution involving the protest paradigm.

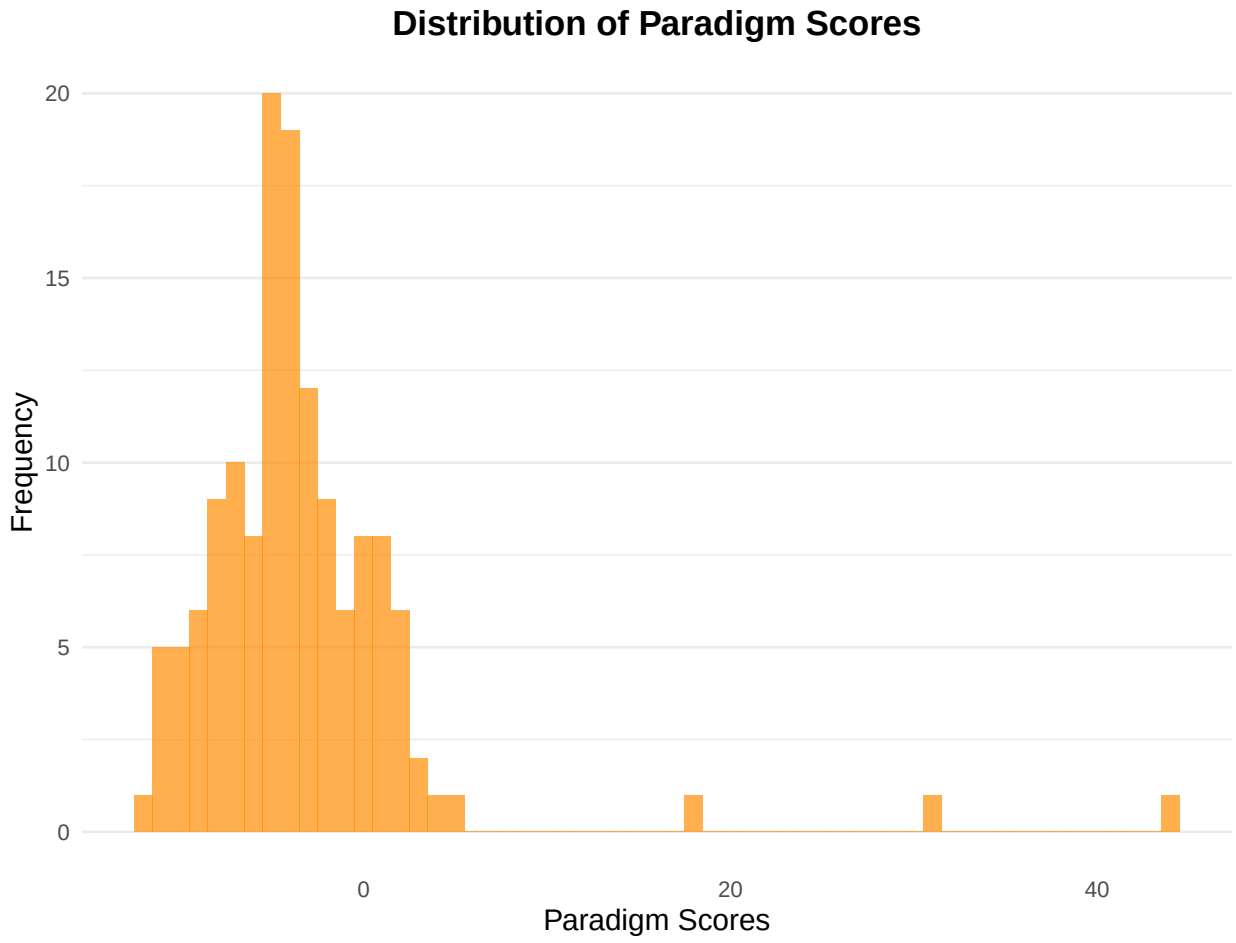

Figure 3: Distribution of paradigm scores derived from the Bradley-Terry model

4.3 Relationship between Sentiment and Paradigm Scales

Armed with these measures, one first-order question of interest is the degree to which the two dimensions are correlated. By calculating the sentiment and paradigm scales separately using independent instances of the LLM, it is possible that underlying associations between sentiment and focus might reveal themselves. Substantively, one might expect that coverage which focuses on the protesters' demands or the reasons why the protest occurred could be more positively valenced. I plot the association between sentiment and paradigm scores in Figure 4, revealing that there is a modest positive association across these two dimensions, but not as high an association as I initially hypothesized.

4.4 Topic Modeling

I also explored these stories and scores using Latent Dirichlet Allocation (LDA) topic modeling. Using the elbow method, I settled on 10 different topics and analyzed each topic's average score. The paradigm scores have a much wider range for the 10 topics, with protests relating to the "state" being the least demands-based and protests surrounding Israeli prime minister Benjamin Netanyahu being the most focused on protest causes. The sentiment scores show less variation across topics compared to the paradigm scores. Most topics

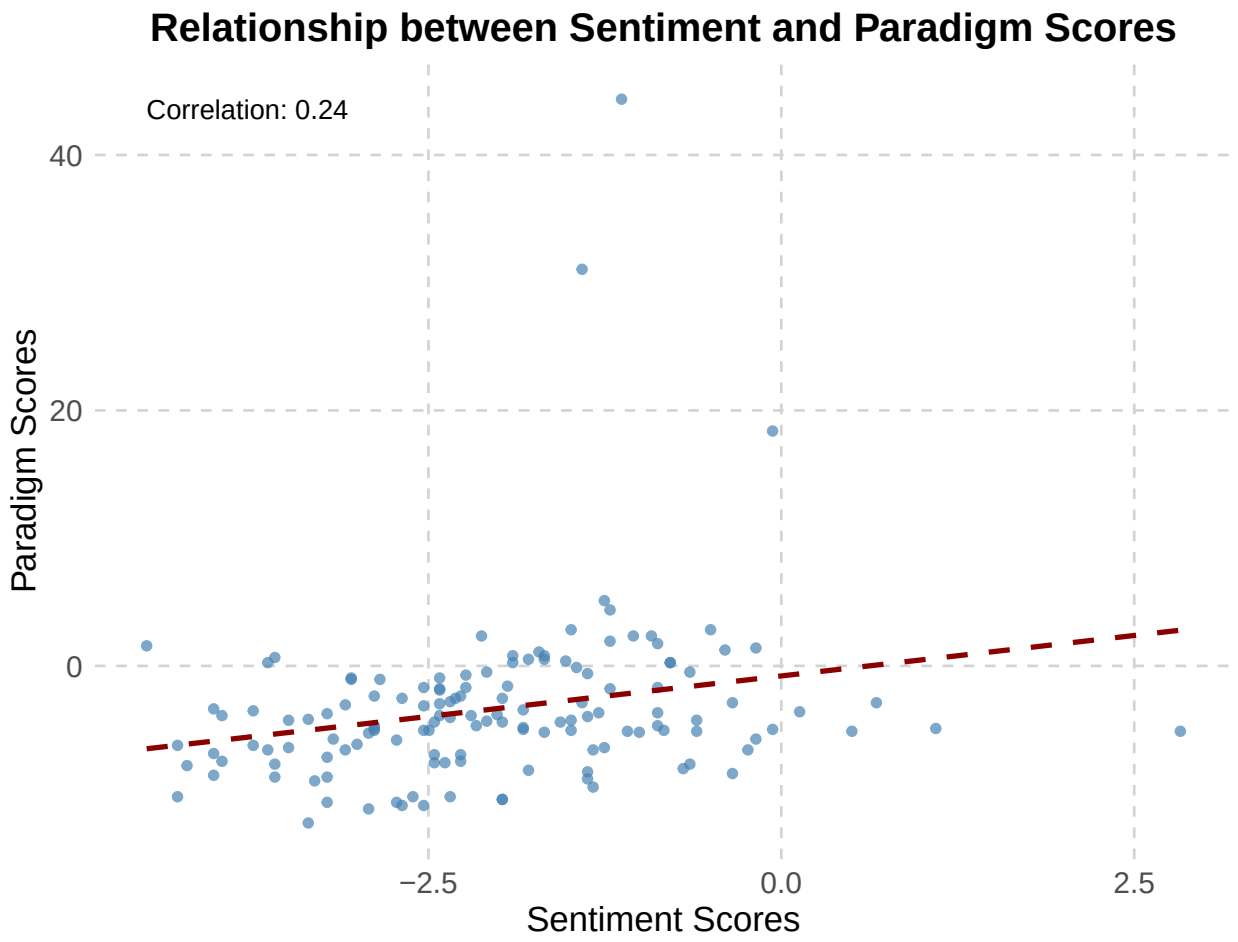

Figure 4: 140 articles arranged by the sentiment score (x-axis, ranging from most negative to most positive coverage) and the paradigm score (y-axis, ranging from most about the tactics used by the protesters to most about the protesters' demands).

have sentiment scores close to the middle of the range, implying a relatively neutral tone. However, articles related to the three keywords "Biden," "President," and "Gaza", as well as the topic surrounding keywords "New," "Police," and "Hundreds" have slightly higher sentiment scores, implying a more positive portrayal of the protests and protesters in these particular contexts.

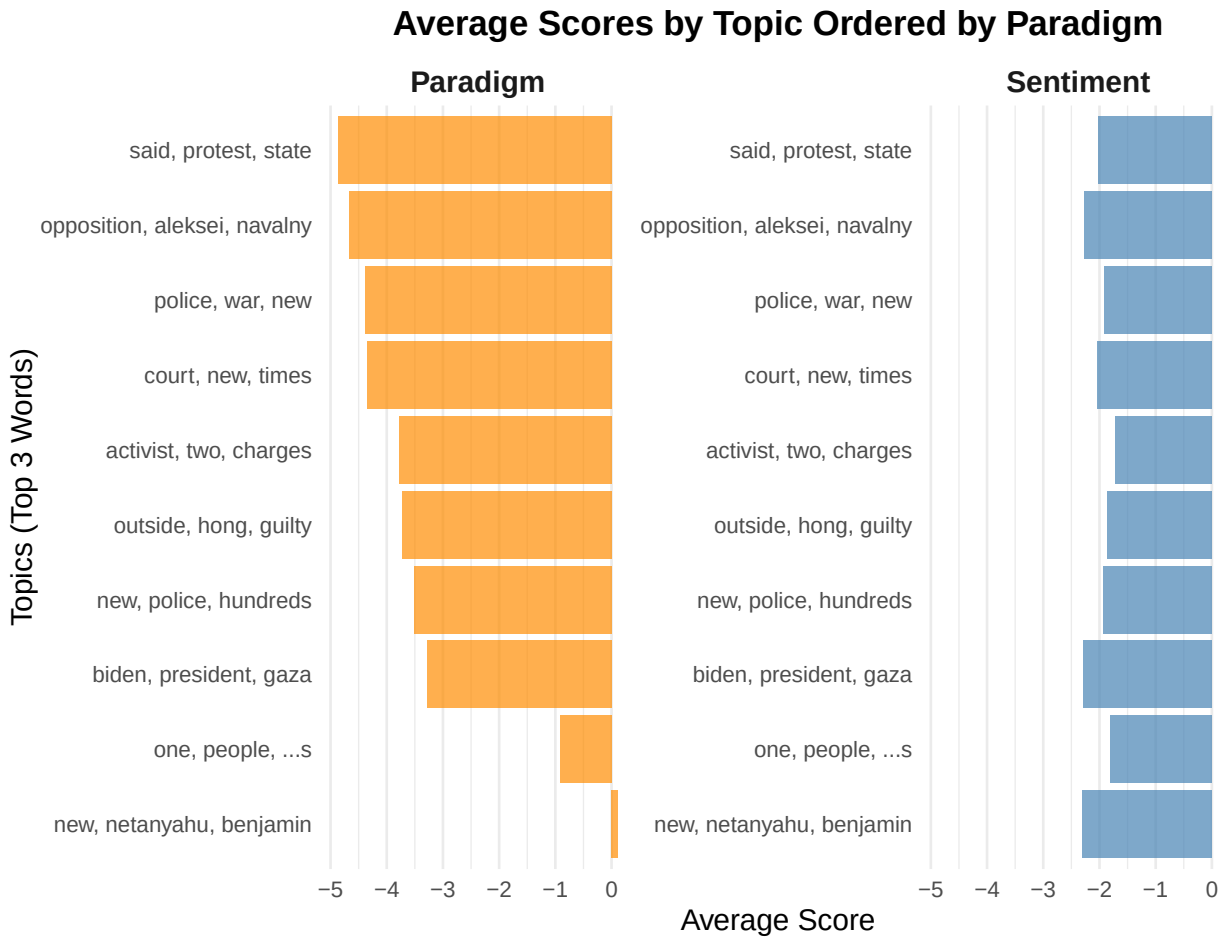

Figure 5: Average paradigm and sentiment scores for each of the 10 topics, characterized by their 3 most frequent words

5 Conclusion and Future Work

By harnessing the analytical capabilities of LLMs, I created continuous, unidimensional scales that were able to capture the latent dimensions of sentiment and paradigm in protest coverage. My findings demonstrate the potential of expanding upon the work of Wu et al. (2023) to analyze how media outlets portray social and political issues, as well as using this particular approach in other types of research. By feeding the LLM the full text to be compared, researchers can apply this method to other comparison types, creating scales for previously difficult to measure metrics like widespread sentiment or classification tasks.

5.1 Limitations

The results themselves are incredibly promising, but the methodology does have its limitations. This type of model does not come with the ability to easily insert new articles, nor can the model be applied to additional articles without being fully re-calibrated. There is no typing tool that can easily incorporate additional data

points, but I hope to explore the idea of using other LLMs as a means of scoring new articles. I hope to use the scores that I've compiled as benchmarks, and then be able to fit new articles within that spectrum, based on whether they are more or less positive than the existing articles, for example. However, this idea remains unexplored and its practicality is unknown.

Additionally, ties within the pairwise comparisons are difficult to deal with. From limited exploration, it seems that the model will simply either reply with "None", meaning that comparison must be dropped from the result set, or the model will simply select the second article it was passed, which would lead to biased data.

In terms of training time and modeling cost, running the pairwise comparisons on OpenAI's platform can be costly and time-consuming, and it is challenging to know a priori how many ties will occur and how much data will be lost as a result. The run-time here for only 9,730 comparisons was in excess of 30 hours, so larger datasets could run into a bottleneck.

5.2 Future Work

Future research should focus on several key areas to address the limitations of this project and expand its scope, and I hope to address most of these points over the summer. First, incorporating more protest data from additional time periods would add stability and interpretability to my models, especially with respect to the topic modeling. Additional sources beyond The New York Times would add to the generalizability of the model and make the pairwise comparisons more comprehensive.

I also hope to use these scores as inputs for new models, such as a regression that uses protest characteristics as the inputs for determining a relationship between protest features and media coverage. This would help validate the measures while also utilizing them in more typical research scenarios.

References

Bradley, Ralph Allan, and Milton E Terry. 1952. "Rank analysis of incomplete block designs: I. The method of paired comparisons." *Biometrika* 39(3/4): 324–345.

Brown, Danielle K., and Summer Harlow. 2019. "Protests, Media Coverage, and a Hierarchy of Social Struggle." *The International Journal of Press/Politics* 24(4): 508–530.

Gottlieb, Julian. 2015. "Protest News Framing Cycle: How The New York Times Covered Occupy Wall Street." *International Journal of Communication* 9(1): 231–253.

Hancock, Braden, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Re. 2018. "Training Classifiers with Natural Language Explanations." *Proc Conf Assoc Comput Linguist Meet* , 1884–1895.

McLeod, Douglas M., and James K. Hertog. 1992. "The manufacture of 'public opinion' by reporters: informal cues for public perceptions of protest groups." *Discourse & Society* 3(3): 259–275.

Miric, Milan, Nan Jia, and Kenneth G. Huang. 2022. "Using supervised machine learning for large-scale classification in management research: The case for identifying artifical intelligence patents." *Strategic Management Journal* 44(2): 491–519.

Weaver, David A., and Joshua M. Scacco. 2013. "Revisiting the Protest Paradigm: The Tea Party as Filtered through Prime-Time Cable News." *The International Journal of Press/Politics* 18(1): 61–84.

Wu, Patrick Y, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. 2023. "Large Language Models Can Be Used to Estimate the Latent Positions of Politicians." *arXiv preprint arXiv:2304.12350* .