Exploring Headlines Using Sentiment Analysis and Word Embedding

Nathaniel Wang
Department of Computer Science
Columbia University
new2127@columbia.edu

Directed Research Report – Fall 2022
Advisor: Prof. John Kender

January 3, 2023

Abstract

News headlines accompany news stories across the majority of news transmission mediums, such as in the form of video titles and article headers. As news headlines tend to contain a condensed description of the contents of the accompanying story, we explore a collection of recent headlines regarding the Russo-Ukrainian War in an attempt to understand the differences in framing used in reporting a common event between cultures. To this end, we apply both frequency-based sentiment analysis techniques as well as word embedding and projection in order to visualize any potential distinctions. Our results suggest that there are measurable differences between Russian and Western news article headlines in terms of sentiment, in addition to apparent differences in framing strategies.

1 Introduction

News headlines serve to provide a potential reader, listener, or watcher of news content with an informationally condensed and intriguing preview of said content. As such, they can be seen as an effective summary of a given news story. Due to this condensed nature, news headlines have adopted a non-standard of writing distinct from conventional written and spoken English. In this study, we hope to determine whether or not American and British news sources use demonstrably different appeal and framing strategies within their headlines than their Russian counterparts via computational means. We apply basic sentiment analysis, as well as attempt an embedding-based approach to a collection of news headlines surrounding the Russo-Ukrainian War from a diverse battery of Russian and American/British sources.

Prior work on framing and news headlines have largely focused on news occurrences that concern only a single country, such as American gun violence [5] or Russian political strategies [3]. Studies that do attend to an international event, such as [2], may also rely primarily on descriptive methods as opposed to quantitative. As such, we believe that this study can lay the groundwork for further studies into quantitative cross-cultural headline framing.

2 Methods

2.1 Ground-Truth

We establish a ground-truth by asking human volunteers to sort a collection of 40 English-language news headlines into two groups of his or her choosing. The volunteers were not made aware that half of the headlines were chosen from a selection of Russian news outlets, with the other half originating from a mix of American and British news agencies. A participant that decided to sort the headlines into "Western" and "Russian" headlines was able to construct two groups that reflected the headlines' origins fairly accurately, with a score of 31 out of 40, compared to the average score of 24.8. This indicates that there likely exists some detectable stylistic difference between headlines depending on the culture of origin.

2.2 Data Collection

The data used in this study consists of 6,416 English-language headlines scraped from the webpages of six news agencies during late autumn of 2022. 3,154 headlines were scraped from a combination of three online Russian news sources (TASS, RT, and The Moscow Times.) Using the same method, 3,262 headlines were collected from two American (The New York Times, Fox News) and one British (The Guardian) counterpart. We aim to use a diversity in news sources, such that any differences found between Russian and Western headlines are less likely to have originated from the stylistic decisions of a single news agency, as opposed to a broader cultural trend.

Western NE	Eastern NE
Zelensky	Russia
Biden	Kremlin
White House	Putin
US	Moscow
Poland	Russians
Ukrainian	Russian
NATO	China
EU	Belarus
UN	Iran
West	
Zelenskiy	
Western	

Table 1: List of important named entities.

2.3 Sentiment Analysis

We perform an analysis of the sentiments surrounding specific named entities present within the headline data, as it is possible that some differences between the headlines would be reflected in the tone in which key figures, events, and locations are described. Named entities of note were determined by hand and are described in Table 1.

We use NLTK [1] to tokenize and identify named entities per headline. Each named entity is then checked for membership in the list of named entities. Should the named entity be relevant, the sentiment score of the headline it belongs to would be associated and logged. The score is determined using the Afinn [6] word list, wherein a positive Afinn score i.e. > 0 would be simplified into a "positive" rating, and a negative score i.e. < 0 would be simplified into a "negative" rating. A "neutral" rating is also used for sentences with an Afinn score of 0. We then compare the sentiment scores per named entity category against either culture of origin, in order to determine if there exist correlations between named entities and the culture writing the headline via sentiment.

2.4 Word Embedding

We also use Doc2Vec [4] to generate 8-dimensional word embeddings for each headline. As embeddings have been demonstrated to accurately reflect word relationships in a geometric context, our rationale is that differences in word use for the purposes of framing the headline should be detectable through analysis of its embedded representation.

We also used K-means clustering via Scikit-learn [7] on the embedded representations as a method of determining similarities in headlines. Similar headlines should generally share the same cluster, given the relatively close distance between their representations.

3 Results

3.1 Sentiment Analysis

Sentiment scores are normalized along the axis of named entities, in order to account for the fact that there were significantly more instances of Eastern named entities, such as "Ukraine." We then compare "positive," "negative," and "neutral" sentiments per category of named entity based on news source culture of origin, as described in Figure 1.

Our sentiment analysis results suggest that there may be a correlation between the culture of origin of a headline and the sentiment associated with certain named entities. Eastern news sources tend to write headlines that reflect more positively on Eastern named entities, while seemingly downplaying negative aspects simultaneously. Similarly, Western news sources may also be more inclined to a negative outlook, as expressed by the higher rates of negative headlines and fewer positive and neutral examples. The lack of neutral examples may also indicate a tendency to skew towards more emotionally charged headline styling.

3.2 Word Embedding

We choose to cluster the data twice: once on two centroids and once on six centroids. We attempt clustering around two centers on the basis that the clusters could form along cultural differences. Conversely, the option of centering around six different means is chosen in the hope that it will provide enough granularity to reveal specific similarities in topic selection and, consequently, framing strategy.

We then project our embedded representations onto two dimensions using PCA and observe the presence of distinct, albeit noisy, clusters. We then backtrace the headlines from each cluster and aggregate the most frequent tokens per cluster, minus determinants, pronouns, and certain conjunctions.

Red	Green	Brown	Yellow	Blue	Orange
ukraine	ukraine	russia	ukraine	ukraine	ukraine
russia	russia	ukraine	russia	ukrainian	russia
russian	russian	nuclear	russian	russian	war
says	war	putin	ukrainian	russia	putin
us	sanctions	russian	war	says	russian
putin	day	war	kherson	nuclear	says
nuclear	invasion	says	putin	plant	are
biden	eu	energy	eu	troops	ukrainian
ukrainian	know	europe	happened	shelling	will
kremlin	ukrainian	is	kyiv	kherson	us

Table 2: 10 most popular tokens per cluster, excluding determinants, pronouns, conjunctions.

The 2-way K-means clusters are generally too vague to offer insights through this methodology. The visualization demonstrates this and can be found in Figure 3. However, the 6-way clusters seem to group along observably different headlines. Table 2 corresponds with the visualization shown in Figure 2 and shows the most common tokens per group. Words such as "Russia" and "Ukraine" appear frequently, as expected, given that they describe the two primary combatants involved in the conflict.

Aside from this, certain clusters seem to feature different key topics more heavily than others. For instance, "biden" makes an appearance in the top 10 tokens for the red cluster, exclusively. Given that this cluster also consists primarily of Russian-sourced news headlines, there is evidence to suggest that Russian news outlets tend to focus on the American president more heavily than their Western counterparts.

Similarly, the blue cluster is populated by headlines that address the battles fought during the course of the conflict. Keywords such as "shelling," "kherson," "troops," and "nuclear" "plant" directly reference battlegrounds and notable associated events, such as a strike on a nuclear power plant. This cluster is also primarily composed of Russian headlines, suggesting that Russian news agencies are also more prone to covering the war's development through battle coverage.

Clusters that primarily feature Western headlines, notably the brown and orange ones, seem to focus on different events. The brown cluster suggests that a more popular framing mechanism among these American and British sources is to cast the war through a domestic economic lens, focusing on "nuclear," "energy," and "europe." What the orange cluster focuses on is less clear from this perspective. However, its heavy use of first-person pronouns (which have been omitted from the table) and state-focused vocabulary (e.g. "are," "will") suggest that these headlines employ tactics to view the war through a more directly personable or empathetic lens.

	Red	Green	Brown	Yellow	Blue	Orange
Western	37%	42%	67%	52%	35%	72%
Russian	63%	58%	33%	48%	65%	28%

Table 3: Proportions of Western/Russian headlines per cluster.

	Red	Green	Brown	Yellow	Blue	Orange	Total
Western	392	495	660	590	348	777	3262
Russian	656	681	326	551	634	306	3154
Total	1048	1176	986	1141	982	1083	6416

Table 4: Total Western/Russian headlines per cluster.

4 Conclusion

The purpose of this study is to broadly compare the framing mechanisms between two dissimilar cultures as pertains to news headlines covering a single, shared event. Using the United States and the United Kingdom as one broad culture and Russia as the foil, we are able to perform a limited analysis on differences in framing strategy by applying sentiment analysis and embedded document representations.

Our results appear to demonstrate that Russian news sources are more likely to speak positively regarding Russian affairs and figures than Western outlets. In the same vein, Western headlines tend to feature more emotional language and a negative outlook than their eastern counterparts.

Our findings also suggest that Russian headlines are more predisposed to using battlefield happenings as framing devices for the war at large, in addition to speaking more frequently on western leadership. By the same token, Western sources seem to favor discussion on the economic impacts of the conflict. It is possible that Western sources favor more personal appeals, as well.

Potential future work that could improve on this study include using more sophisticated sentiment classification techniques, such as a neural network-based approach. Work which marries the two analytical approaches explored here in order to determine which frames are more likely to carry positive or negative emotional valences (e.g. discussing battlefield events with a positive spin) would also have the potential provide further insight into the problem explored here.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[2] Sumayya Ebrahim. The corona chronicles: Framing analysis of online news headlines of the COVID-19 pandemic in italy, USA and south africa. *Health SA Gesondheid*, 27:1683, February 2022.

[3] Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1393. URL `https://aclanthology.org/D18-1393`.

[4] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents, 2014. URL `https://arxiv.org/abs/1405.4053`.

[5] Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1047. URL `https://aclanthology.org/K19-1047`.

[6] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, May 2011. URL `http://ceur-ws.org/Vol-718/paper_16.pdf`.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

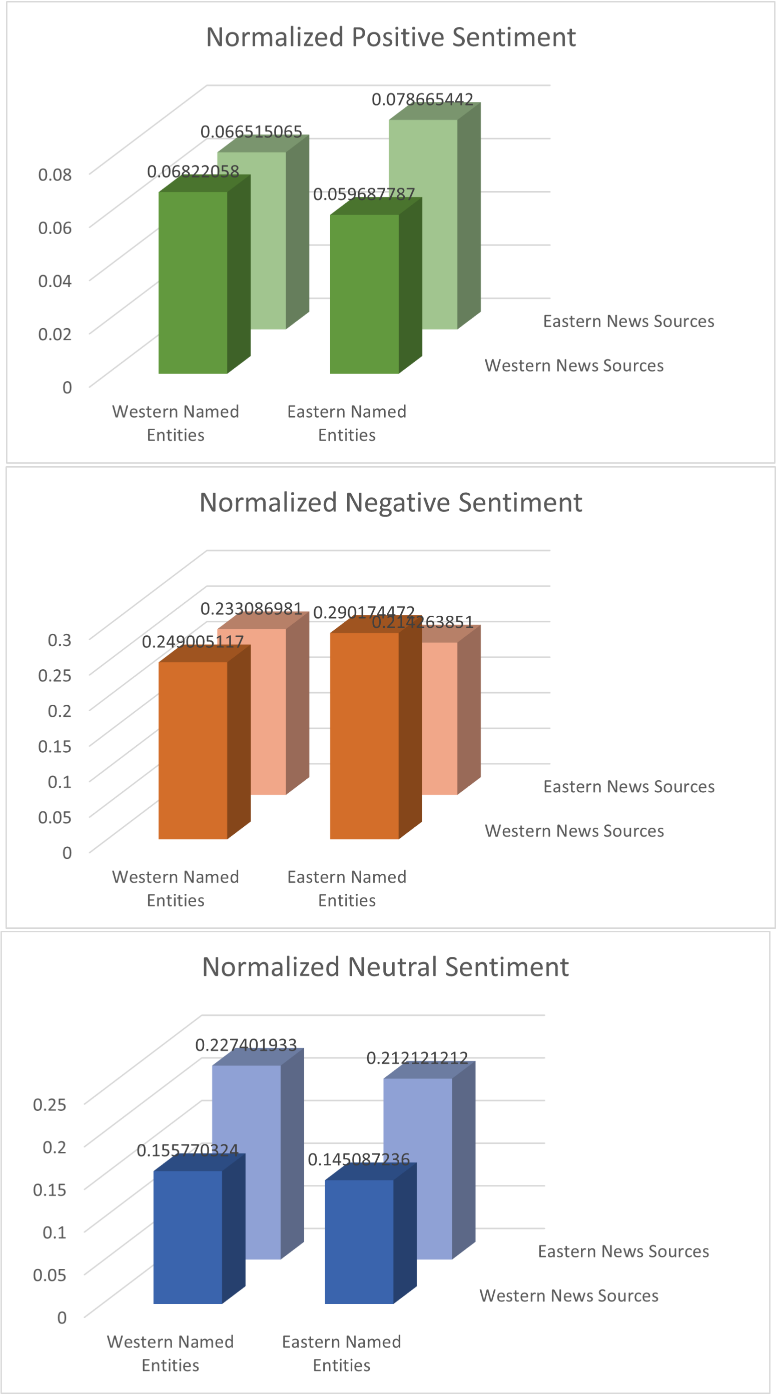

Figure 1: Sentiment scores between named entities and news sources.

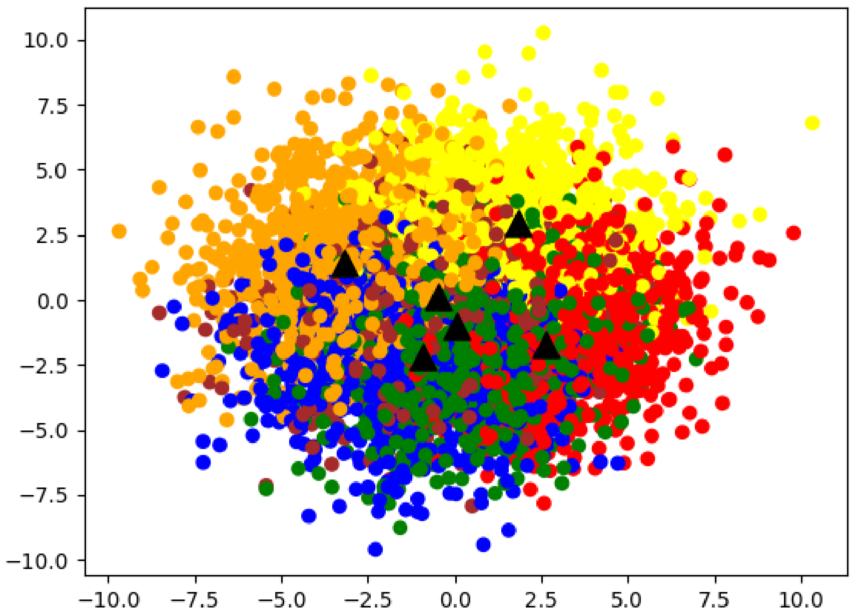

Figure 2: PCA plot for 8-dimensional headline embeddings. Colors denote 6 clusters, trianges denote centroids.

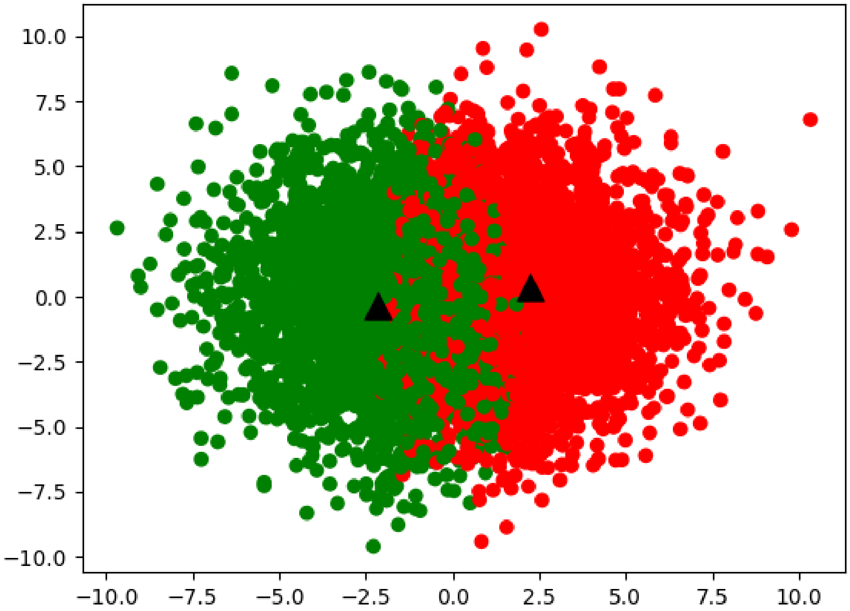

Figure 3: PCA plot for 8-dimensional headline embeddings. Colors denote 2 clusters, trianges denote centroids.