Understanding the Impact of Geographical Bias on News Sentiment: A Case Study on London and Rio Olympics

Swati
swati@ijs.si
Jožef Stefan Institute,
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute,
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

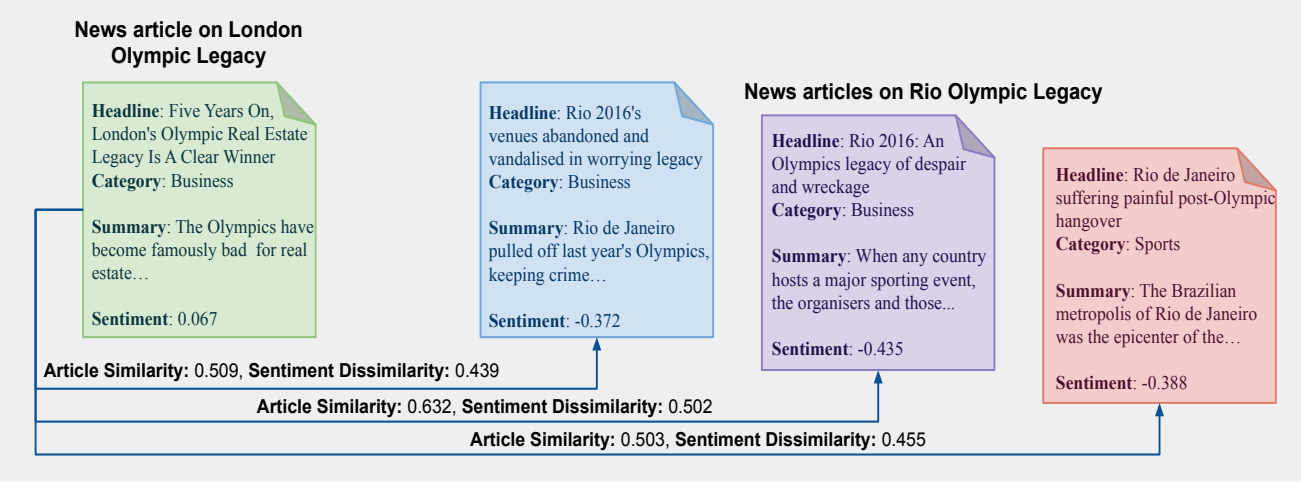

Figure 1: An example to illustrate the impact of geographical location on the sentiment of similar news articles.

ABSTRACT

There are various types of news bias, most of which play an important role in manipulating public perceptions of any event. Researchers frequently question the role of geographical location in attributing such biases. To that end, we intend to investigate the impact of geographical bias on news sentiments in related articles. As our case study, we use news articles collected from the Event Registry over two years about the Olympic legacy in London and Rio. Our experimental analysis reveals that geographical boundaries do have an impact on news sentiment.

KEYWORDS

Bias, News Bias, Geographical Bias, Olympics, Semantic Similarity, Sentiment Analysis, Dataset

1 INTRODUCTION

Claims of bias in news coverage raise questions about the role of geography in shaping public perceptions of similar events. Based on the geographical location, multiple factors, such as political affiliation, editorial independence, etc., can influence the way news articles are generated. Although it is well known that biased news can have more influence on people's thinking and decision-making processes [7, 9], it is nearly impossible to produce an article without any bias. Biased news articles have the potential to induce a variety of political and social implications, both direct and indirect. For instance, any political controversy presented from a specific perspective may alter the voting pattern [4, 1, 6].

There are different forms of news bias, and geographical bias is one of them. It exists if the sentiment polarity of similar articles published in different geographical location is contradictory or varies significantly. Sentiment analysis methods, which are commonly used to determine news bias [3, 14], can be used to examine the shift in sentiment polarity in similar news articles. Now, an intriguing question arises: Is geographical bias a factor affecting news sentiment? This study seeks to answer the above question by identifying and comparing sentiments of similar news articles. In doing so, we demonstrate how geographical location impacts the sentiments of similar articles. We also investigate this impact in relation to several news categories such as politics, business, sports, and so on.

The Olympic Games are a symbol of the greatest sports events in the world. Every edition leaves a number of legacies for the Olympic Movement, as well as unforgettable memories for each host city, whether positive or negative. In this regard, we select news articles about the Olympic legacy in London and Rio as a case study for our analysis.

We use Event Registry[1] [10] to collect English news articles, along with their sentiment and categories, published between January 2017 and December 2020. We use the popular Sentence-BERT (SBERT) [12] embedding to represent the articles and then compute the cosine similarity between them to identify similar article pairs.

Our data and code can be found in the GitHub repository at *https://github.com/Swati17293/geographical-bias*.

Information Society 2021, 4 October 2021, Ljubljana, Slovenia

[1] https://eventregistry.org

1.1 Contributions

The paper's contributions are as follows:

- We propose a task of analyzing the impact of geographical bias on the sentiment of news articles with data on the Olympic legacies of Rio and London as a case study.
- We present a dataset of English news articles customized to the above-mentioned task.
- We present experimental results to demonstrate the aforementioned impact of geographical bias.

2 RELATED WORK

The Majority of the sentiment analysis methods for news bias analysis depend on the sentiment words that are explicitly stated. SentiWordNet[2], which is a publicly available lexical resource used by the researchers for opinion mining to identify the sentiment inducing words that classify them as positive, negative, or neutral.

Melo et al. [5] collected and analyzed articles from Brazil's news media and social media to understand the country's response to the COVID-19 pandemic. They proposed using an enhanced topic model and sentiment analysis method to tackle this task. They identified and applied the main themes under consideration in order to comprehend how their sentiments changed over time. They discovered that certain elements in both media reflected negative attitudes toward political issues.

Quijote et al. [11] used SentiWordNet along with the Inverse Reinforcement Model to analyze the bias present in the news article and to determine whether the outlets are biased or not. The lexicons were first scored for the experiments using SentiWordNet and then fed to the Inverse Reinforcement model as input. To determine the news bias, the model measured the deviation and controversy scores of the articles. The findings lead to the inference that articles from major news outlets in the Philippines are not biased, excluding those from the Manila Times.

Bharathi and Geetha [3] classified the articles published by the UK, US, and India median as positive, negative, or neutral using the content sentiment algorithm [2]. The sentiment scores of the opinion words and their polarities were used as input to the algorithm.

Existing research investigates news bias using sentiment analysis methods, but, unlike our work, it does not provide a suitable automated method for analyzing the impact of geographical bias on news sentiment.

3 DATA DESCRIPTION

3.1 Raw Data Source

We use **Event Registry** [10] as our raw data source which monitors, gathers, and delivers news articles from all around the world. It also annotates articles with numerous metadata such as a unique identifier for article identification, categories to which it may belong, geographical location, sentiment, and so on. Its large-scale coverage can therefore be used effectively to assess the impact of geographical bias on news sentiment.

3.2 Dataset

To generate our dataset, we use a similar data collection process as described in [13]. Using the Event Registry API, we collect all English-language news articles about the Olympic legacy in London and Rio published between January 2017 and December 2020. We consider an article to be about the Olympic Legacy in London/Rio if the headline and/or summary of the article contains the keywords '*London*'/'*Rio*', '*Olympic*', and '*Legacy*'.

For each article, we then extract the summary, category, and sentiment. The article summaries vary in length from 290 to 6,553 words. Sentiment scores ranges from −1 to 1. We select seven major news categories, namely business, politics, technology, environment, health, sports, and arts-and-entertainment, and remove the rest of the categories. After excluding the duplicate articles we end up with 8,690 and 5,120 articles about the Olympic legacy in London and Rio respectively.

4 MATERIALS AND METHODS

4.1 Methodology

The primary task is to compute the average difference in sentiment scores between similar news articles about the Olympic legacies in Rio and London. The stated task can be subdivided and mathematically formulated as follows:

(1) Generate two distinct sets of news articles A_1 and A_2, one about the London Olympic legacy and the other about the Rio Olympic legacy. For each $a_i \in A_1$ find a list of $a'_j \in A_2$, where a_i is the i^{th} article in set $A_1 = \{(a_1, s_1), (a_2, s_2)...(a_n, s_n)\}$ and a'_j is the j^{th} article in set $A_2 = \{(a'_1, s'_1), (a'_2, s'_2)...(a'_m, s'_m)\}$ which is the closest match (c.f. Section 4.1.1) to a_i. Here, $n = |A_1|$ and $m = |A_2|$.
(2) For each list, calculate D_{ij} to represents the difference between the sentiment scores s_i and s'_j of the articles a_i and a'_j.
(3) Calculate the average difference D of sentiment scores.
(4) Calculate the percentage of similar article pairs with reversed polarity and those with unchanged polarity.

The secondary task is to assess the primary task with respect to news categories, i.e. to calculate the average difference D of sentiment scores for similar articles in each category.

In the following subsections, we discuss the tasks mentioned above in greater detail.

4.1.1 Article Similarity. We embed the articles in sets A_1 and A_2 to construct sets $F_1 = \{f_1, f_2...f_m\}$ and $F_2 = \{f'_1, f'_2...f'_n\}$. While alternative embedding approaches can be utilized, in this study we select the popular Sentence-BERT (SBERT) [12] embedding to extract 768-dimensional feature vectors to represent the individual articles in F_1 and F_2.

For each article a_i in A_1, we compute the similarity score[3] between a_i and every article a_j in A_2 using the cosine similarity metric $Sim^{cos}(a_i, a'_j)$ (Eq 1). We consider articles a_i and a'_j to be similar only if their similarity score is greater than 0.5.

$$Sim^{cos}(a_i, a'_j) = \frac{f_i \cdot f'_j}{||f_i||\ ||f'_j||} \tag{1}$$

where f_i and f'_j represents the embedded feature vectors of article a_i and a'_j.

The similarity score ranges from −1 to 1, where −1 indicates that the articles are completely unrelated and 1 indicates that they are identical, and in-between scores indicate partial similarity or dissimilarity.

4.1.2 Average Sentiment Dissimilarity. For every pair of similar articles a_i and a'_j, we calculate the difference D_{ij} between their sentiment scores s_i and s'_j. To calculate the average sentiment

[2] http://sentiwordnet.isti.cnr.it/

[3] https://en.wikipedia.org/wiki/Cosine_similarity

Table 1: Category-wise confusion matrix to show the percentage of similar article pairs with respect to their sentiment polarity.

	Sports		Business		Politics		Environment		Health		Technology		Arts & Entertainment	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Pos	77	10	62	28	42	18	55	18	29	12	87	4	59	16
Neg	11	2	7	4	23	16	14	12	12	46	1	0	7	18

Table 2: Confusion matrix to show the percentage of similar article pairs with respect to their sentiment polarity.

	Positive	Negative
Positive	69	15
Negative	11	4

Table 3: Distribution of average sentiment difference across news categories for similar article pairs with identical category.

News category	Average Sentiment Difference
Sports	0.19
Business	0.20
Politics	0.18
Health	0.16
Environment	0.22
Technology	0.14
Arts and Entertainment	0.19

dissimilarity score D, we add all D_{ij} and divide it by the total number of similar article pairs.

5 RESULTS AND ANALYSIS

In our experiments, we compare 44,492,800 possible article pairs for similarity and discover 375,008 similar pairs. The comparison in terms of sentiment similarity reveals that if two articles from different geographical regions are similar, in our case Rio and London, the average difference in their sentiment scores is 0.171. In addition, as defined in Table 2, we calculate the percentage of similar article pairs based on their sentiment polarity. It's worth noting that the polarity of the article is completely reversed 27% of the time, indicating the impact of geographic region on sentiments.

It is because the success of mega-events such as the Olympics in a particular host city is heavily influenced by its residents' trust and support for the government [8]. It can be viewed positively as a national event with social and economic benefits, or negatively as a source of money waste. While the Olympics have left an economic and social legacy in London, a series of structural investment demands in Rio raise the question of whether or not the Olympics was worthwhile for the entire country.

5.1 Impact of news categories

The impact of news categories on the sentiments of similar articles with identical categories from different geographical regions is shown in Table 3. It demonstrates that certain news categories have a greater impact than others. Figure 2 depicts this distinction more clearly.

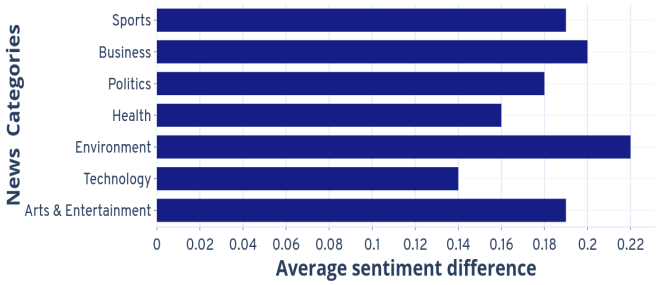

Figure 2: Distribution of average sentiment differences across categories for similar articles in the same category.

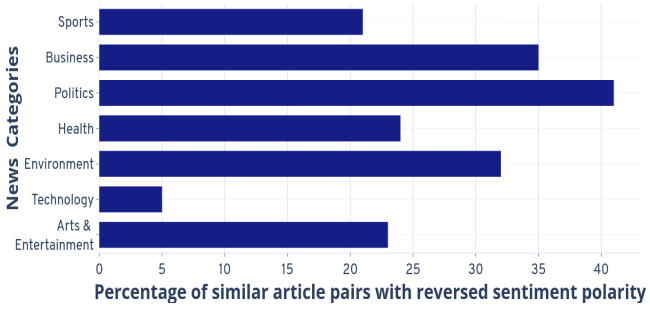

Figure 3: An illustration of the effect of category on sentiment polarity.

The categorical distribution of the percentage of similar article pairs in terms of sentiment polarity is shown in Table 1. *'Politics'* has the highest percentage of articles with reversed polarity, while *'technology'* has the lowest. Categories such as *'business'* and *'entertainment'*, though not as clearly as *'politics'*, exhibit the same bias.

This disparity arises from the fact that, in contrast to other categories, politics is most influenced by geographical boundaries, whereas science and technology are typically location independent. Since politics has such a large influence on shaping beliefs and public perceptions, it is frequently twisted to fit a particular narrative of a story. It is inherently linked to geographical borders, and it can be extremely polarizing depending on the geographical region.

6 CONCLUSIONS AND FUTURE WORK

In this work, we use news articles about the Olympic Legacy in London and Rio as a case study to understand how geographical boundaries interplay with news sentiments.

We begin by presenting a dataset of news articles collected over two years using the Event Registry API. We compute the cosine similarity scores of all possible embedded article pairs, one

from each set of Olympic legacy articles (London and Rio). We use the popular Sentence-BERT for article embedding and then compute the sentiment difference between similar article pairs. From 44,492,800 possible article pairs we end up with 375,008 similar pairs.

In our analysis, we discovered that the sentiment reflected in similar articles from different geographical regions differed significantly. We also investigate this difference in relation to different news categories such as politics, business, sports, and so on. We find a significant difference in news sentiment across geographical boundaries when it comes to political news, while in the case of news in technology, the difference is much smaller. We find that articles in categories such as politics and business can be heavily influenced by geographical location, articles in categories such as science and technology are typically location independent.

In the future, we plan to identify the most frequently mentioned topics in the Olympic legacy corpus to see how they affect the news sentiment of articles about different geographical locations. Since our study is limited to English news articles, we intend to learn more about the role of cultures and languages in this bias analysis. We also intend to broaden our investigation to discover the adjectives used to describe the negative and positive legacies of Rio and London. Such an analysis would aid in understanding the expectations from cities such as Rio (the first in South America to host the Olympics) in comparison to London.

7 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Dan Bernhardt, Stefan Krasa, and Mattias Polborn. 2008. Political polarization and the electoral effects of media bias. *Journal of Public Economics*, 92, 5-6, 1092–1104.

[2] Shri Bharathi and Angelina Geetha. 2017. Sentiment analysis for effective stock market prediction. *International Journal of Intelligent Engineering and Systems*, 10, 3, 146–153.

[3] SV Shri Bharathi and Angelina Geetha. 2019. Determination of news biasedness using content sentiment analysis algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, 16, 2, 882–889.

[4] Chun-Fang Chiang and Brian Knight. 2011. Media bias and influence: evidence from newspaper endorsements. *The Review of economic studies*, 78, 3, 795–820.

[5] Tiago de Melo and Carlos MS Figueiredo. 2021. Comparing news articles and tweets about covid-19 in brazil: sentiment analysis and topic modeling approach. *JMIR Public Health and Surveillance*, 7, 2, e24585.

[6] Claes H De Vreese. 2005. News framing: theory and typology. *Information Design Journal & Document Design*, 13, 1.

[7] John Duggan and Cesar Martinelli. 2011. A spatial theory of media slant and voter choice. *The Review of Economic Studies*, 78, 2, 640–666.

[8] Dogan Gursoy and KW Kendall. 2006. Hosting mega events: modeling locals' support. *Annals of tourism research*, 33, 3, 603–623.

[9] Daniel Kahneman and Amos Tversky. 2013. Choices, values, and frames. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 269–278.

[10] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

[11] TA Quijote, AD Zamoras, and A Ceniza. 2019. Bias detection in philippine political news articles using sentiwordnet and inverse reinforcement model. In *IOP Conference Series: Materials Science and Engineering* number 1. Volume 482. IOP Publishing, 012036.

[12] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, (November 2019). http://arxiv.org/abs/1908.10084.

[13] Swati, Tomaž Erjavec, and Dunja Mladenić. 2020. Eveout: reproducible event dataset for studying and analyzing the complex event-outlet relationship.

[14] Taylor Thomsen. 2018. Do media companies drive bias? using sentiment analysis to measure media bias in newspaper tweets.