Visualizing Polarity-based Stances of News Websites

Masaharu Yoshioka
Hokkaido University
Sapporo-shi, Hokkaido, Japan
yoshioka@ist.hokudai.ac.jp

Myungha Jang James Allan
UMass Amherst
Amherst, MA, USA
{mhjang, allan}@cs.umass.edu

Noriko Kando
National Institute of Informatics (NII)
Chiyoda-ku, Tokyo, Japan
kando@nii.ac.jp

CEUR-WS.org/Vol-2079/paper2.pdf

Abstract

We develop a novel framework that helps identify potential bias in news websites to support users who are exposed to news articles with a wide variety of political leanings. We propose a *polarity-based stance* (PS), a vector that represents how often a website publishes articles that are positive or negative with regard to a topic. We derive PS using the GDELT database and visualize the news websites' stances. We demonstrate the utility of our framework via a case study of the 2016 US Presidential Election.

1 Introduction

There are two types of users when it comes to their pattern of news navigation. The first type already has particular news websites that they trust and actively use by accessing them directly for news. Such websites tend to demonstrate the same political stances or leanings as their users. As a result, the articles that they read are likely ones that already share their ideologies. The other type, those who are less politically engaged, use a news aggregation website that shows a compiled list of news articles from various sources. A key difference in the two approaches is that the latter type of user, because content is the primary factor in selecting articles, is exposed to news from more diverse sources, which demonstrates a wider array of political stances. Users must therefore use their own judgment to selectively digest what they read, especially for controversial topics.

Many users judge the trustworthiness of new websites based on their political bias. Hence, we propose a novel framework that represents the bias of news websites toward a particular topic as a vector. Using this framework, we then visualize stances of news websites toward a given topic. For this, we define a *polarity-based stance*, a vector that represents bias toward a particular topic of a website using the polarity of stances. This allows us to visualize the stance of news websites, guiding users for the potential bias of the articles published by the websites. We demonstrate the usefulness of our framework via the case study of 2016 US President Election using the GDELT database[1].

2 Polarity-based Stances

We formally define a polarity-based stance, $\overrightarrow{PS}_w$, as a two-dimensional vector that denotes the stance of a website w. We first assume that each article of the website has one of three stances: positive, negative, or neutral. We let $\overrightarrow{PS}_w = [p, n]$ where p is the ratio of positively-stanced articles and n is the ratio of negatively-stanced articles for a particular topic. Note that the stance has been identified beforehand. We discuss how to use the GDELT database to derive this vector.

Copyright © 2018 for the individual papers by the papers' authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors.

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

[1] https://www.gdeltproject.org

2.1 Dataset

The GDELT database is one of the largest news article repositories collected by the Google Jigsaw project. It is a useful resource for multifaceted analysis for news articles because it has a large amount of data and contains the metadata including the source website that are automatically extracted from various NLP algorithms for the crawled articles [YK16].

We use *tone*, one type of automatically generated metadata, to derive $\overrightarrow{PS}_w$. Tone refers to the average attitude of the article, which is computed by the difference between the percentage of positive and negative terms in the document[Pro15]. Calculation of polarity score based on the term matching is simple and it is better to use more sophisticated methodology [RR15]. However, due to the large numbers of the articles for analysis, it is almost impossible for the GDELT users to crawl the all text of the articles and calculate scores for them. For the case study analysis later, we use articles from the GDELT database published on the 2016 US Presidential Election during a three month period that includes voting day (see Table 1).

Table 1: The description on the article dataset in the GDELT database used.

Period	Sep 1, 2016 - Nov 30, 2016
# of Articles	22.4M (0.2M per a day)
# of News Websites	44,624

2.2 Deriving Polarity-based Stances

We compute $\overrightarrow{PS}_w$ using the tone score provided by the GDELT database. Let d be a news article published by a news website w and t be the tone of d. We classify the document stance s_d into one of three classes: positive (1), neutral (0), and negative (-1). The stance is derived from t given a threshold σ using the equation

$$s_d = \begin{cases} 1 & t > \sigma \\ 0 & -\sigma < t < \sigma \\ -1 & t < \sigma \end{cases} \quad (1)$$

We then define a polarity-based stance ($\overrightarrow{PS}_w$) for a website (w) using the equation

$$\overrightarrow{PS}_w(\tau) = \Big(\frac{\sum_{d \in w_\tau} (\mathbb{1}[s_d = 1])}{|w|}, \frac{\sum_{d \in w_\tau} (\mathbb{1}[s_d = -1])}{|w|} \Big) \quad (2)$$

where w_τ is a set of articles on τ published by w. By plotting these stances on a graph, users can compare stances of different news websites.

In addition, bias can be identified by comparing stances of the similar topics or one with a particular topic and general topic.

3 Case Study

We demonstrate the utility of our approach via a case study of the 2016 US Presidential Election around two topics: Donald Trump and Hillary Clinton. To visualize the polarity-based stances for these topics, we estimate the set of news articles on each topic using a simple Boolean query. When an article references both Trump and Clinton, there is ambiguity about which topic is indicated by the tone. We therefore identify the set of articles that exclusively references only one of the topic to compute the polarity-based stances (see Table 2).

Table 2: The numbers of articles for the boolean queries of "Donald Trump"(DT) and "Hillary Clinton"(HC) (The numbers in the parenthesis indicates the total number of articles that contain DT and HC)

Query	# of articles
DT - HC	677,307 (1,516,225)
HC - DT	388,162 (1,227,080)
DT or HC	838,918

Table 3 shows distributions of tone (-100 to 100) in the articles retrieved by DT-HC and HC-DT as queries using their number. For both queries, numbers of articles for negative tone are larger than one for positive, but the difference is not so large in general[2]. So we set the value of $\sigma = 1$ in equation 1 for this experiment. However, it is better to check how σ affects the final results in the future research.

Table 3: Distribution of tone (using number of articles) in the retrieved articles

Tone	DT-HC	HC - DT
$[-100, -3]$	188,709	89,283
$(-3, -2]$	95,665	51,781
$(-2, -1]$	109,575	67,006
$(-1, 0]$	123,231	74,878
$(0, 1)$	65,554	42,080
$[1, 2)$	46,999	31,618
$[2, 3)$	23,528	15,510
$[3, 100]$	24,046	16,006

Figure 1 and 2 show the scatter plot of polarity-based stances of various news websites for the Trump and Clinton topics. In these plots, we include news websites that published more than 30 articles for the particular topic. Each circle indicates a news website with a radius that signifies the number of articles. The top 20 news websites that published the most articles exclusively on Trump and Clinton are indicated by colored circles. Note that a new website with a small number of articles is shown as a point.

To visualize the bias of the websites (toward Trump or Clinton), we plot the absolute difference of positive

[2]Most of the articles have their tone values between -3 to 3 (DT-HC:69%, HC-DT:73%)

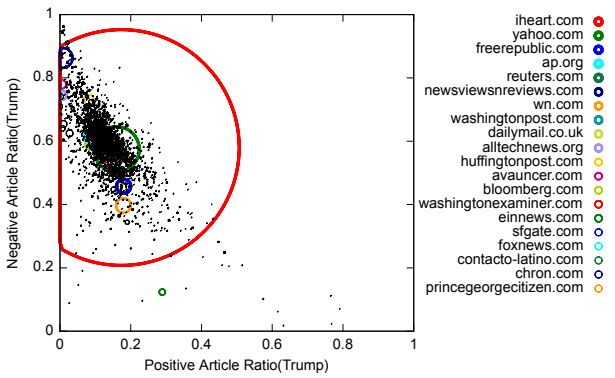

Figure 1: The polarity-based stances of the Trump topic visualized in a scatter plot

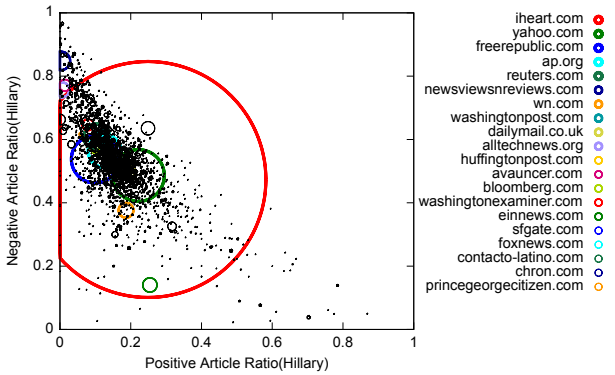

Figure 2: The polarity-based stances of the Clinton topic visualized in a scatter plot

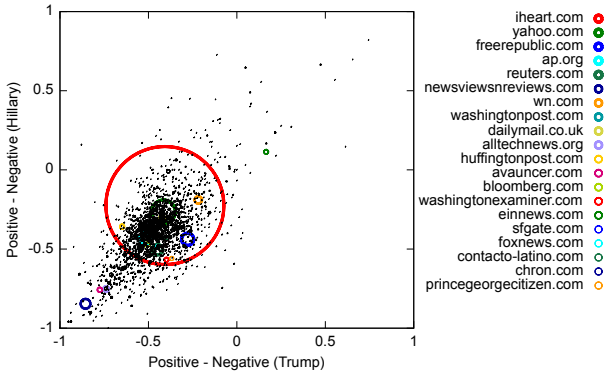

Figure 3: Diff(Trump) and Diff(Clinton) to compare their polarity-based stances

and negative articles ratio for Trump and Clinton in Figure 3. We let $\textsc{Diff}(\tau)$ to be the absolute difference between the two components of $\overrightarrow{PS}_w(\tau)$. We plot $\textsc{Diff}(Trump)$ and $\textsc{Diff}(Clinton)$ for comparison (See Figure 3). The websites whose bias towards the two topics are the same are plotted on the line of $(\textsc{Diff}(Trump) = \textsc{Diff}(Clinton))$. The points at the top left of the plot are the articles that are positively-stanced towards Clinton, and the ones at the bottom right are positively-stanced towards Trump. The plot helps us identify the news websites whose polarity-based stances are completely different between the two topics. For example, `thebostonpilot.com` has (0.15, 0.27) for "Trump", and (0.16, 0.65) for "Clinton" and `sci-tech-today.com` as (0.02, 0.90) for "Trump", and (0.41, 0.25) for "Clinton". It is important to take into account such bias when such big difference happens.

4 Conclusion

In this paper, we propose a framework to visualize stances in the dimensions of polarity of news websites to identify a potential bias in the articles that are published by them. We define a vector named Polarity-based Stance and demonstrate the utility via a case study of 2016 U.S. Presidential Eleciton, and that the GDELT database is a useful resource for this type of analysis. As a future work, we plan to apply our framework to a variety of topics for evaluation. We observe that some topics generally have a higher positive, or negative articles than the others. We plan to study how to take this factor into account to visualize stances in an useful way.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 16H01756.

References

[Pro15] GDELT Project. The gdelt global knowledge graph (gkg) data format codebook v2.1, 2015.

[RR15] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14 – 46, 2015.

[YK16] Masaharu Yoshioka and Noriko Kando. Comparative analysis of gdelt data using the news site contrast system. In *The first International Workshop on Recent Trends in News Information Retrieval (NewsIR)*, 2016.