What's in the News? Towards Identification of Bias by Commission, Omission, and Source Selection (COSS)

Anastasia Zhukova[1,2], Terry Ruas[2], Felix Hamborg[3], Karsten Donnay[4], Bela Gipp[2]
[1]University of Wuppertal, Germany
[2]University of Göttingen, Germany
[3]Heidelberg Academy of Sciences and Humanities, Germany
[4]University of Zurich, Switzerland
zhukova@gipplab.org

ABSTRACT

In a world overwhelmed with news, determining which information comes from reliable sources or how neutral is the reported information in the news articles poses a challenge to news readers. In this paper, we propose a methodology for automatically identifying bias by commission, omission, and source selection (COSS) as a joint three-fold objective, as opposed to the previous work separately addressing these types of bias. In a pipeline concept, we describe the goals and tasks of its steps toward bias identification and provide an example of a visualization that leverages the extracted features and patterns of text reuse.

KEYWORDS

new analysis, media bias, text alignment, text reuse, paraphrase identification

1 INTRODUCTION AND RELATED WORK

The literature on media bias has found that editorial choices in the news production process, such as bias by commission and omission of information and source selection (COSS), strongly affect public perceptions [1]. This finding is particularly alarming since today's news production system faces pressure to minimize reporting costs [4, 12]. Consequently, journalists often rely on the same news source, copy reports, or the (factual) information in other reports [5]. This phenomenon, also termed pack journalism, tends to lead to a lower quality of reporting, as journalists fail to independently verify the information they report [11].

Unlike scientific publications, where sources of information must be documented explicitly, news articles typically contain no citations [2]. However, much of the information in articles typically originates from previously published articles, newswire reports, or press releases [6, 12]. Compared to its sources, which information is included or excluded in an article is typically opaque to the news reader [8]. Especially when information is reused as paraphrases where different to the original source wording is used, which eventually leads to biased reporting [10].

In the past two decades, computer science has addressed the problem of automated identification of text reuse. Studies of text reuse substantially focus on (1) plagiarism detection, semantic textual similarity, and paraphrase identification, (2) text reuse in journalism, blog posts, and newswire reports, on the Web, and in Wikipedia, (3) information flow analysis, story diffusion and propagation, and news story chains, (4) novelty detection. However, the adaption and application of the automated analysis approach to the news domain and the study of bias are only just emerging [8].

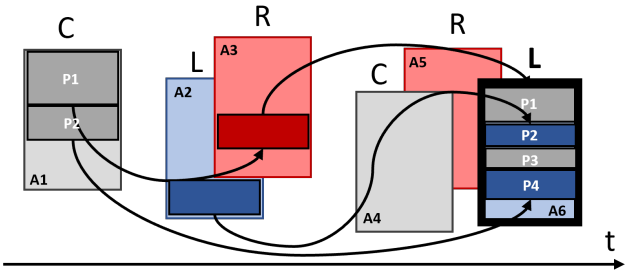

Figure 1: An example of visualizing the extracted information by pipeline for identifying bias by COSS. The identified text reuse in a set of related articles is ordered by date-time. The figure shows that a seed article A6 with left polarity (L) contains both original and paragraphs with reused information. For example, paragraph P3 contains original information labeled as center-oriented (C), although the article belongs to a left-oriented publisher. On the contrary, paragraph P4 was reused from article A1 and has changed its polarity from the original central to the left.

Previous approaches considered identifying bias by commission and omission as two separate tasks [3, 9]. They used statistical approaches that relied on direct text reuse, e.g., TFIDF, and focused only on one type of bias. Moreover, the existing independent news aggregators that cover a full political spectrum, such as AllSides[1], Ground.News[2], and The True Story[3] either focus on one of these types of biases or perform the analysis on an article level. In this paper, we propose a methodology and a concept pipeline that identifies bias by COSS as a three-fold objective. We address text reuse on a more conceptual level, such as paraphrasing, and aim at identifying text reuse on a higher level of granularity, e.g., paragraphs.

2 PIPELINE CONCEPT

The pipeline for identification of bias by COSS has two purposes: (1) analysis of a given seed article against a collection of event-related articles to identify which parts of it are original and which are reused, (2) identification of patterns in information flows in

[1]https://www.allsides.com/
[2]https://ground.news/
[3]https://thetruestory.news/

a collection of event-related articles to explore a bigger picture on information reuse. Both of the flows require the same pipeline stages: (a) candidate retrieval, (b) source retrieval and text alignment, (c) construction of a graph of text reuse, (d) pattern analysis, (e) visualization of the extracted information.

Candidate retrieval obtains articles reporting the same event. The step extracts event-related documents from a large database of news articles, e.g., LexisNexis[4], CommonCrawl[5], MediaCloud[6], and The GDELT Project[7]. To retrieve related documents, the system should use either an event-descriptive query and a time frame for this event or a seed document with its timestamp. Alternatively, for the system evaluation in a closed environment, candidate retrieval supports reading a provided set of related articles that contain all required attributes, e.g., a timestamp. Similar to Ground.News, to each article, we assign a polarity label induced from an outlet, e.g., each article from Fox News will be labeled as "R" for right or conservative slant.

Source retrieval and text alignment are the core steps in identifying information reuse that analyze which parts of text are reused from which source(s). Unlike most existing methods for text alignment that identify copy-pastes or word permutations, we focus on identifying paraphrased sentences or paragraphs that convey the same message but use different and possibly loaded wording.

Polarity classification enables revisiting both original and reused paragraphs and checks if the outlet-inferred labels correspond to the labels from a polarity classifier. Such a polarity relabeling tracks how the same message evolves over time and across the outlets (see Figure 1). Training a reliable classifier requires a large balanced dataset to incorporate variance of the biased language [7].

A graph of text reuse stores articles, their paragraphs, and both extracted and assigned attributes to enable exploration of the patterns of information reuse. The relations between the paragraphs encode the strength of semantic similarity between the paragraphs, and the time codes of the articles enforce a directed graph, which is required for source identification.

Statistical and network analysis aims at identifying **patterns** of text reuse that may induce bias by COSS. Analysis of the article's origin includes determining how many paragraphs originate from which sources with which polarities. For example, if an article reuses a significant amount of information from neutral sources, such as news agencies, we could conclude that this article is reliable and unlikely bias-prone. On the contrary, if an article consists of too many slanted paragraphs with no sources, it might indicate a lack of trustworthiness in this article. Analyzing the information for bias by the commission includes analysis of which information with which polarity tends to be reused, how often and how reused paragraphs change polarity, how long information continues being reused after the original publishing, etc. On the contrary, analyzing patterns of bias by omission includes identifying which parts of the source articles were not picked up by which articles were excluded from discussions. For example, if a source article is left-oriented, a right-oriented article could reuse only excerpts that report about an event itself but omit parts that fall into the liberal agenda.

[4]https://www.lexisnexis.com/
[5]https://commoncrawl.org/
[6]https://mediacloud.org/
[7]https://www.gdeltproject.org/

Visualization is an efficient way to enable researchers and news readers to explore the extracted information in a tempo-oriented graph structure (see Figure 1). Additionally, visualization depicts such identified features as the strength of semantic similarities between paragraphs [9], the original and assigned polarity of paragraphs and articles, and highlighted patterns that may indicate biases of each of the three types.

3 CONCLUSION

In this paper, we propose a concept of identification of bias by COSS as a joint three-fold objective. Compared to the previous systems identifying these types of bias via simple direct text reuse, our system leverages a solid research basis in text plagiarism detection and recent advances in paraphrase identification with semantic similarity. Revealed cases of paraphrasing help determine the source of articles and how reused information changed over time and news articles.

REFERENCES

[1] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[2] Darrell Christian, Paula Froke, Sally Jacobsen, and David Minthorn. 2014. *The Associated Press stylebook and briefing on media law.* The Associated Press.

[3] Jonas Ehrhardt, Timo Spinde, Ali Vardasbi, and Felix Hamborg. 2021. Omission of Information: Identifying Political Slant via an Analysis of Co-occurring Entities. In *Information between Data and Knowledge.* Schriften zur Informationswissenschaft, Vol. 74. Werner Hülsbusch, Glückstadt, 80–93. https://epub.uni-regensburg.de/44939/

[4] Susanne Fengler and Stephan Ruß-Mohl. 2008. Journalists and the information-attention markets: Towards an economic theory of journalism. *Journalism* 9, 6 (2008), 667–690.

[5] Russell Frank. 2003. These crowded circumstances' when pack journalists bash pack journalism. *Journalism* 4, 4 (2003), 441–458.

[6] Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough, and Scott Piao. 2001. The METER corpus: a corpus for analysing journalistic text reuse. In *Proceedings of the corpus linguistics 2001 conference*, Vol. 1. Citeseer.

[7] Lukas Gebhard and Felix Hamborg. 2020. The POLUSA dataset: 0.9 M political news articles balanced by time and outlet popularity. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020.* 467–468.

[8] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20, 4 (2019), 391–415.

[9] Felix Hamborg, Philipp Meschenmoser, Moritz Schubotz, Philipp Scharpf, and Bela Gipp. 2021. NewsDeps: Visualizing the Origin of Information in News Articles. *Wahrheit und Fake im postfaktisch-digitalen Zeitalter: Distinktionen in den Geistes-und IT-Wissenschaften* (2021), 151–166.

[10] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated Identification of Media Bias by Word Choice and Labeling in News Articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL).* 196–205.

[11] Jonathan Matusitz and Gerald-Mark Breen. 2012. An examination of pack journalism as a form of groupthink: A theoretical and qualitative analysis. *Journal of Human Behavior in the Social Environment* 22, 7 (2012), 896–915.

[12] Zizi Papacharissi and Maria de Fatima Oliveira. 2008. News frames terrorism: A comparative analysis of frames employed in terrorism coverage in US and UK newspapers. *The international journal of press/politics* 13, 1 (2008), 52–74.