Chair of Data Science
Prof. Michael Granitzer

UNIVERSITÄT PASSAU
Fakultät für Informatik und Mathematik

Research Proposal

Student Name: Long Vu
Doctoral Supervisor: Anastasia Zhukova
Date of this Document: xx.07.2021

Towards identification of bias by commission and source selection in a collection of news articles by identification of reused text fragments

1 INTRODUCTION

Media bias occurs during the news production process and it plays a part in our everyday life. Nevertheless, rather than reflecting the perspective of individual journalists or articles, the term means a pervasive or widespread bias contrary to the journalism standards. In various countries, there is controversy over the extent and direction of media bias [5].

There are many types of media bias that may occur at any step in the news production process. One of them is where journalists have to include sources into their articles, i.e press releases, news articles, researches reports to increase the trustworthiness of their articles, however, they may consult sources supporting one perspective, this type is called bias by source selection [5]. Journalists can ultimately choose which information from the sources to include and which to exclude in the article to be written, this yields another type of bias called bias by commission and omission [5].

Several critics contend that the media is capable of slanting news stories to favor particular groups, interests, or groups of people using a variety of manipulative techniques. Bias is often subtler and thus harder to detect than intentional lying and censorship, which are at their most blatant. In contrast, many journalists insist that true objectivity cannot exist and that biased media should not be viewed negatively since those biases can indeed influence the presentation of facts. When journalists believe there will always be some bias, they aim to eliminate it as much as possible rather than to limit it, and they attempt to present a balanced view [5].

There are several projects which aim to mitigate media bias, either by providing changes in grammar or vocabulary choice in the article or provide multiple ways to have a full picture of an event. NewsCube [1] is an Internet news service aiming at mitigating the effect of media bias, by automatically creating and providing and promptly provides readers with multiple classified viewpoints on a news event of interest. As such, it effectively helps readers understand a fact from plural of viewpoints and formulate their own, more balanced viewpoints.

Media bias may appear in many different areas, such as politics, arts, technology, business. The US-political area will be discussed mainly in this project. Typically political spectrum in real life is often divided by left wing and right wing, representing Democracy Party and Republican Party [15]. However, the magnitude of political media bias is usually extended as *Far Left, Left, Center, Neutral, Right, and Far Right* as some media outlets tend to not lean or partly lean to one side of the spectrum, this type of rating is used by AllSides [12] and GroundNews [10].

One of the core systems for these kinds of service to run is their text similarity system, which is responsible for clustering the articles written on the same event(s). The key idea behind these models is document vectorization - a technique that translates a single word into a multi-dimensional vector of real numbers which allows easy calculating the of similarity between them. There are many models which empower this technique, an example is the vector space model TFIDF, which is easy to compute, but can't capture the semantic of the sentence, to neural network word2vec, their performance is already evaluated in many studies.

Shahmirzadi et al. [4] conducted a research in 2020 that compares different methods, in particular, TFIDF and its 2 more extensions, latent semantic indexing, and doc2vec, comes to a conclusion, that TFIDF performs very well in most cases, whereas complex embedding methods are only useful if the text is very condensed and the similarity detection is relatively coarse. Extensions of TFIDF, such as adding n-gram do not seem to have better performance.

Another approach in studying text similarities is Cross Document Coreference Resolution [7] (CDCR), which is a subarea of from Coreference Resolution (CR) by not only processing text within a single document, but in a collection of documents. CDCR extracts and resolves mentions into entities/events, which refer to as real entities, this way CDCR can help identify related terms and articles, which helps in detecting multiple viewpoints of a particular event.

Methods from plagiarism detection (PD) - which is a field of computer science with the aim of identifying unauthorized information reuse in documents [5] - can also be applied to detect bias by source selection and by commission, e.g using methods such as simple text-matching to find 1:1 duplicates [17] or identify paraphrases on the paragraph levels. However, research on bias by source selection using PD methods is not active in recent times, most likely because authorized copyediting is a component in the news production process.

Bias by commission and omission is created in a similar way as to bias by source selection and did not receive much interest in computer science as well. Most researches applied on this type was conducted by social scientists. However, although existing computer science approaches haven't been applied to identified this type of bias, yet they are considered as potential candidates. For example, aspect-level browsing which is implemented in the news aggregator NewsCube, enables users to view different perspectives on political topics [1]. Centering resonance analysis (CRA) - works by constructing a graph with each node representing a term that is contained in the noun phrases of a given text [1] - is also employed by social scientists as well.

2 Goal of the Project/Thesis

2.1 Problem Statement

As mentioned previously, bias by *source selection* and by *commission and omission* are the 3 forms of media bias. Although there are existing methods in computer science that are considered the potential for the analysis of those forms. Yet they have not drawn much attention from computer scientists, and hence leave open questions unanswered - which will be discussed in the next section.

In this project, I would like to investigate and implement a new system for automated analysis of these forms of bias by utilizing existing methodologies to make a step towards bias identification.

2.2 Research Objective(s)

The new approach for analysis of bias by source selection and commission and omission is expected to help us solve the following problems:

2.2.1 Common stories detection

Before going into the problem's details, we need to understand what exactly is a "story"? In the scope of this project, a *story* is a sentence that describes comprehensively a news article that writes about a real-world event, the event can happen and finish in the past, or can be currently occurring, e.g school shooting, presidential election. Here the term *event* will be used alongside the term *story* interchangeably. Therefore, stories detection means, by feeding collections of corpus to that we would want group them together by which story they are covering to see is happening/has happened in the real world.

In particular, the system has to summarize every article in the form of a short sentence, then detect and cluster similar topics together, and create a sentence that best describes all those articles. By grouping them together, in the later phases we can analyze bias between related articles only and therefore avoid unnecessary computations.

2.2.2 Narrowly event-related issue identification

Journalists often follow the 4Ws or sometimes 5Ws questions [19], which are the questions whose answers are considered basic information gathering. The questions typically are *Who, What, When Where, Why*, hence produce articles that may cover multiple subtopics defined by specific events, location, time, and participant(s). Following this phenomenon, we call *narrowly event-related articles* are the articles that cover the same story, they may or may not contains other subtopics.

The main goal is to extract these related subtopics within these articles, the result will allow us to see how a specific story develops, e.g when and where does the story originates, or answer questions, does the story evolve into another story.

2.2.3 Original source article identification

A narrowly event-related story can be reported in multiple articles as described above, but everything in the article may not be written by the author. Instead, the author may copy in whole or in parts from other articles. And as mentioned earlier, this will cause bias by source selection.

By detecting the origin of a story and the source of each article, we can further detect if an article is slanted by how the author chooses source articles for reference.

2.2.4 Detection of the similarities between text fragments

Bias by commission and omission have the same effects as bias by source selection. The Media Insight Project's research [18] has concluded that more than 80% of related news articles add no new information and only reuse information contained in previously published articles. This type of bias, too does not draw much interest from scientific researches, hence leave open questions unanswered, such as how reliable a piece of information reported in an article?

Here the system is expected to be able to identify the fragments that are potentially taken from other articles, this will help us see how much attention a story is taken, how it develops in the agenda, and fades, or detect bias by commission and omission.

2.3 Expected Results

In this project, a news dataset, preferable POLUSA (more on Section 3), will be fed to the system and should yield a list of possible topics, alongside their stories (or subtopics). There will be a list of articles covering each subtopic, and one of them will be the potential source article. You can see a visual representation of this hierarchical structure in Figure 1.

Also, the project should provide us with a table indicating how paragraphs from one article are reused by another article, here the measurement can be the similarity score or the number of reused paragraphs or can be both.

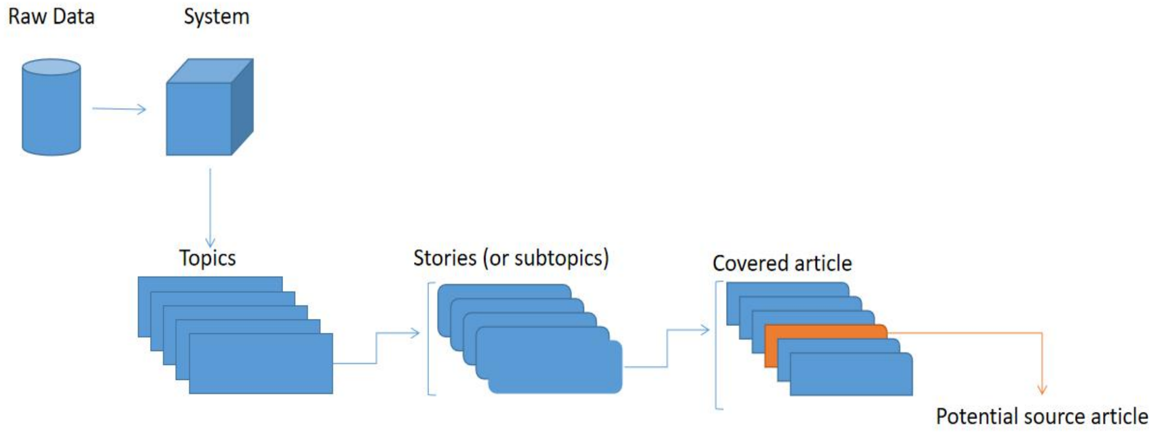

Figure 1: The project results in a list of topics, inside each of each are stories (or subtopics). Each subtopic contains a collection of articles written about it, where one of them can be the potential source article (colored with orange).

3 METHODOLOGY

The project will be divided into multiple phases based on the research objectives.

As the first step, I will focus on the task *event detection* (or topic modeling), where I will utilize LDA as a baseline method to extract most appeared terms within the articles and cluster them using common similarities distance methods, such as Cosine similarity score [31]. Here I will use the dataset POLUSA - a big dataset that represents the online media landscape as perceived by an average US news consumer [2], the dataset contains nearly a million articles and covers a lot of diverse topics, hence making it suitable for my project.

In the 2nd phase where the narrow subtopics are to be extracted. I will employ *Giveme5W1H* - an open-source library developed by Hamborg et al. [22] which is capable of extracts entities based on the journalistic 5W and 1H questions and summarizing into the main theme of an article - to train a model on a dataset published by the creator themselves, then use the model to extract the main events from every article from POLUSA, then vectorize and cluster similar ones. I will employ TFIDF as a baseline method for vectorizing the events since Hamborg et al. [33] conducted a similar research. The model will be trained to classify subtopics with NaiveBayes or K-Nearest Neighbors on the ECB+ corpus [24] - a dataset of news articles organized on topic and subtopic level. In order to define an event, only 4Ws (excluded the Why and How) entities will be taken into consideration for defining an event, since the other types are more into sentiment entities. The system's result will be evaluated by calculating the accuracy value against the dataset.

The next phase is where the originality of an article is evaluated. Here I'm planning to follow the NewsBERT [25] approach, which is based on BERT [29] but can distill knowledge from pre-trained language models in order to work on news datasets more efficiently. I will train the model on PAN14-Source-Retrieval dataset [26] since the dataset is trusted for evaluation of different approaches in plagiarism detection competition (PAN 2015 [27]).

In the final phase, I will apply a *paragraph2vec* [32], an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, to build embedding vectors on a dataset specialized for text alignment task, here I will also employ another PAN dataset - the PAN14-Text-Alignment dataset [28] for its usage on PAN competition on plagiarism detection - to identify possible plagiarized paragraphs and cluster similar vectors across articles.

Work Plan

The rough work plan for this project looks as follows, the plan takes roughly 4 months to complete:

Data	Milestone
xx.xx.2022	Start of the project
xx.xx.2022	Sketch a prototype of the project: structuring inputs and outputs of each step, design a workflow of the program
xx.xx.2022	Gather and pre-process dataset(s).
xx.xx.2022	Implement 1st task: *event detection* with LDA and Cosine similarity
xx.xx.2022	Implement 2nd task: sub-topics identification based on the 4W journalistic questions using TFIDF and NaiveBayes or KNN
xx.xx.2022	Implement 3rd task: prepare the NewsBERT model and fine tune with the PAN14 dataset and implement it to identify sources of articles
xx.xx.2022	Implement 4th task: build embedding vectors with paragraph2vec and detect potential reused paragraphs.
xx.xx.2022	Putting all together into a system which will consume articles and produces analyses.
xx.xx.2022	Record results and writing the thesis
xx.xx.2022	Thesis submission

4 References

[1] Park, S., Kang, S., Chung, S., Song, J. 2009. NewsCube: Delivering multiple aspects of news to mitigate media bias

[2] Gebhard, L., Hamborg, F. 2020. The POLUSA Dataset: 0.9M Political News Articles Balanced by Time and Outlet Popularity

[3] Hamborg, F., Meuschke, N., Aizawa, Akiko., Gipp, B. 2017. Identification and Analysis of Media Bias in News Articles

[4] Shahmirzadi, O., Lugowski, A., Younge, K. 2020. Text Similarity in Vector Space Models: A Comparative Study

[5] Hamborg, F., Donnay, K., Gipp, B. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review

[6] Zhou, D., Guo, L., He, Y. 2018. Neural Storyline Extraction Model for Storyline Generation from News Articles

[7] Keshtkaran, A., Yuhaniz, S., Ibrahim, S. 2017. An Overview of Cross-Document Coreference Resolution

[8] Vargas, J. 2017. Narrative Information Extraction with Non-Linear Natural Language Processing Pipelines

[9] Zhou, D., Xu, H., Dai, X., He, Y. 2016. Unsupervised Storyline Extraction from News Articles

[10] Ground News: https://about.ground.news

[12] Nobias: https://nobias.com/

[12] https://www.allsides.com/media-bias/media-bias-ratings

[13] Gentzkow, M., Shapiro, J. 2010. What drives media slant? Evidence from U.S. Daily Newspapers

[14] Media Outlets Bias Rating: https://www.allsides.com/media-bias/media-bias-ratings

[15] Left-right political spectrum: https://en.wikipedia.org/wiki/Left-right_political_spectrum

[16] Hamborg, F. 2020 Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling

[17] NLP Exploration: Topic Models: https://dougfenstermacher.com/blog/nlp-exploration-topic-models

[18] Gipp, B. 2014 Citation-based plagiarism detection.

[19] The Personal News Cycle: How Americans choose to get their news: https://www.americanpressinstitute.org/publications/reports/survey-research/personal-news-cycle/

[20] The Five Ws: https://en.wikipedia.org/wiki/Five_Ws

[21] Zhukova, A. et al., 2020 Interpretable Topic Modeling Using Near-Identity Cross-Document Coreference Resolution

[22] Hamborg, F. et al., 2019 Giveme5W1H: A Universal System for Extracting Main Events from News Articles

[23] Hamborg, F. et al. 2018. Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions. Proceedings of the iConference 2018 (Sheffield, UK, 2018), 355–356

[24] The ECB+ corpus: http://www.newsreader-project.eu/results/data/the-ecb-corpus/

[25] Wu, C. et al. 2021. NewsBERT: Distilling Pre-trained Language Model for Intelligent News Application

[26] PAN14 Originality: Source Retrieval: https://zenodo.org/record/3716010#.Ya1vlJHMKUk

[27] Hagen, M. et al. 2015. Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches

[28] PAN14 Originality: Text Alignment https://zenodo.org/record/3716022#.Ya1wlJHMKUk

[29] Potthast, M. et al. 2014. Overview of the 6th International Competition on Plagiarism Detection

[30] Devlin, J. et al. 2018 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[31] Cosine similarity: https://en.wikipedia.org/wiki/Cosine_similarity

[32] Quoc, Le. Et al. 2014. Distributed Representations of Sentences and Documents

[33] Felix, H. et al. 2021. Omission of Information: Identifying Political Slant via an Analysis of Co-occuring Entities