KU LEUVEN

FACULTY OF
ENGINEERING SCIENCE

Artile-level Media Bias Classification

Stefan Liemawan Adji

Thesis submitted for the degree of
Master of Science in
Electrical Engineering, option
Electronics and Chip Design

Supervisor:
Prof. Myriam de Lhoneaux

Assessors:
Ir. Kn. Owsmuch
K. Nowsrest

Assistant-supervisors:
Ir. An Assistent
A. Friend

Academic year 2023 – 2024

Contents

Abstract

In today's digital, information-rich society, **media** bias poses a significant challenge to the objectivity and credibility of news reporting. As someone living in our current society, one has inevitably encountered some form of bias in the media, either consciously or unconsciously. Media bias can shape our perceptions, influence our opinions, and affect our understanding on various issues. It is crucial to recognise and address this bias to ensure a well-informed and balanced perspective. By being aware of the inherent biases in media sources, individuals can critically evaluate the information they consume and seek out diverse viewpoints to form a more comprehensive understanding of the world. In this paper...

Chapter 1

Introduction

1.1 Overview

This thesis project is part of an Advanced Master of Artificial Intelligence programme—Speech and Language Technology, in collaboration with the Media Bias Group [11], who provided the topic and additional guidance along the project.

1.2 Motivation

Media misinformation and manipulation are rampant in today's landscape. Despite being a major societal issue, there has been hardly enough resources and work dedicated to in the realm of media bias and its broader context.

Considerable work has been done on fake news and its detection, media bias operates on a higher level and cannot be described the same simply fake news. Fake news involves intentionally spreading false information, whereas media bias refers to the distortion or manipulation of information by media outlets, which may or may not be intentional, to favor certain perspectives or agendas. Unlike fake news, media bias can influence public perception subtly through selective reporting, framing, or sourcing, making it a complex and challenging issue to address comprehensively.

To combat media bias effectively, it is crucial to develop robust methodologies and technologies for detecting and analyzing biased content across various media platforms. Additionally, raising awareness about media literacy and critical thinking skills can empower individuals to identify and navigate biased information effectively in today's media landscape. Ultimately, addressing media bias requires concerted efforts from researchers, policymakers, media organizations, and the public to promote transparency, accountability, and integrity in news reporting and consumption.

1.3 Goals

To aid in the work of media bias, the goal of this project are defined as follows:

1. Build a reliable, annotated dataset comprised of articles content and metadata.

2. Propose a method to effectively represent articles in a space vector.
3. Design a system that is able to effectively detect and classify bias on the article level of granularity.
4. Validate the resulting system and dataset.

1.4 About the Media Bias Group

The Media Bias Group [11] was established in mid-2020 by Timo Spinde during his pursuit of a Ph.D. in computer science, having been integrated in the topic since his undergraduate studies, with a vision to aid others perceive news in a more balanced and conscious manner. After a year of planning how a system could uncover bias on a vast scale encompassing millions of articles, he founded the group and forged connections with various partners, particularly those relevant to specific aspects of the project. In just one year, the project has garnered support from multiple other research groups, with around twenty students from seven countries joining to contribute to the system. Since 2021, the group has also begun offering its first Ph.D. positions.

The group is comprised of a collective of scholars across various fields such as Psychology, Linguistics, and Computer Science, with a shared goal to comprehend the factors influencing human perception of news content as biased or one-sided. Currently, the network includes of six main researches and coordinators, twenty one professors and postdocs, as well as eight active students. Numerous publications related to media bias have been published through the network into major conferences such as EMNLP 2021 [26], along with dataset and benchmark creations.

Chapter 2

Media Bias

2.1 Definition

Allsides [14] defines media bias as "The tendency of news media to report in a way that reinforces a viewpoint, worldview, preference, political ideology, corporate or financial interests, moral framework, or policy inclination, instead of reporting in an objective way (simply describing the facts)". This phenomenon has existed and researched since the 1950s (cite), highlighting its enduring presence and impact on public perception. Media bias can manifest in various forms, including the selection of stories, framing of issues, and choice of language. This project would mainly focus on text-level context bias: phrasing bias, spin bias, and statement bias, as described in [25]: statement bias refers to “members of the media interjecting their own opinions into the text”, phrasing bias is characterized by inflammatory words, i.e., non-neutral language, spin bias describes a form of bias introduced either by leaving out necessary information.

While biases may not always be intentional, they might cause significant consequences, possibly leading to inequalities and injustices. Some news outlets tend to use catchy headlines which trick readers into clicking, known as "clickbait", which are often ambiguous or misleading. Biased information can and has been used as a way to shape and influence public opinion [1]. A survey of journalists from United States, Great Britain, Germany, Italy, and Sweden, where they found evidence that journalists’ personal beliefs substantially influence their news decisions, expressed within the stories they choose and the statements they write [23].

2.2 Application

Another particularly dangerous example of media bias is its application in relation to elections. According to a survey done by [2], fake news propagated in social media played a pivotal role in the eventual election of President Trump during the 2016 election. Panagopoulos’ study [21] revealed systemic biases leaning towards Democratic candidates during national and state levels pre-election polls conducted during the 2020 U.S. general election cycle. Rafail et al. [24] examined 201,678 media documents from Tea Party organizations, Fox News, MSNBC, and 785 newspapers, revealing significant differences in how the Tea Party frames itself compared to how other media sources frame the movement, MSNBC portrays it as the worst aspect of the Republican Party, while Fox News sees it as the best, sharply in contrasts with how activists frame the movement as conservative but not strictly Republican, often clashing with Republican Party goals. An analysis by Pew Research Center [15] found that miscalculations observed in the 2020 U.S. election polls would adjust public opinion on issues by an average of just under 1 percentage point. Although errors of such scale would not have produced substantial difference on the American public opinions, this shows the underlying bias within polls specifically and the failure of accurately representing surveys.

The phenomenon of media bias in media has not gone unnoticed, particularly by readers and consumers, lowering public trust on media outlets in general. In the United States particularly, Trust on news media is at an all time low, falling consistently and significantly over the past 20 years [10, 9, 19]. Approximately half of Americans believe that the media is significantly responsible for the political divisions within the United States, with a growing number of Americans losing faith in the media's objectivity and perceive it as actively engaging in ideological wars [9]. Reuters Institute reported only less than half of their respondents (40%) generally trust the majority of news sources, with the US ranked as second-lowest by country (32%) [19].

Readers themselves are not exempt from bias, as they are known to prefer to pick, follow, and consume articles that align with their own beliefs and ideology, an issue known as filter bubble [13]. This can be dangerous as it reinforces existing biases and limits exposure to diverse perspectives, creating echo chambers that hinder critical thinking and informed decision-making. The combination of media bias and filter bubbles can distort reality, perpetuate misinformation, and deepen societal divisions. It is essential for readers to seek out a variety of sources and viewpoints to gain a more balanced and comprehensive understanding of the issues at hand.

2.3 Goal

Ideally, unbiased media contents that objectively and fairly represents multiple or a range of perspectives are desirable, news sources should remain neutral and let readers build their own opinions on the subject [20]. However, this is often unachievable due to human capabilities and resources limitations; journalists cannot possibly possess complete knowledge on every topic, be physically present everywhere, or interview every relevant individual on a significant subject. [14]. Truth and journalism objectivity is a complex matter full of choices and dilemma, where ultimately it falls on the journalist own preferences and criteria [3].

Therefore, instead of eliminating media bias, our goal should be to draw attention to its existence, giving readers awareness of such content [25], ultimately building a tool to defend readers from media manipulation, to let readers know on the quality of the article, and if they fall under a victim of political agenda or indoctrination.

Therefore, the primary objective of a media bias classifier should be to develop an automated system capable of classifying media bias at the article level. This system should be able to recognise and categories bias in news articles from diverse sources, ensuring that readers are aware of what they are reading and therefore can make informed choices and opinions. The classification should encompass various dimensions of bias, including political orientation (e.g., left, right, center), sensationalism, and framing techniques.

Chapter 3

Literature Review

3.1 Dataset

Available datasets for article-level media bias classification are quite limited, often in different formats or covering different types of bias:

- The BASIL dataset [7] (300 articles) is the most commonly used bias dataset out there, containing 300 articles with both article-level and phrase-level annotations. This dataset focuses on informational bias in news articles as it appears more frequently than lexical bias. However, the obvious drawback of this dataset is the low amount of articles included as it is not nearly enough data to get a good working detection model.

- NLPCSS [5] (6964 rows), annotated via Ad Fontes labels, contains article-level textual content with three bias labels (bias, neutral, or unknown). Focusing on political bias and unfairness, this can technically be a decent dataset for work for an article-level classification task.

- The BAT [27] dataset (6345 rows, Ad Fontes labels) is the most suitable for article-level bias detection as it supplies both bias-score and reliability-score for the whole article. However, the dataset itself does not supply textual content of the articles and therefore need to be extended, something that I have been working on as well during my Master's Thesis this and last year.

Annotating bias is not a straightforward task. Traditionally, this is done by hiring experts and journalists to manually read and determine how biased the content is. As with [26], annotations are generally compiled and majority voted to achieve the final annotation given a particular text. It is important to use multiple annotators to minimise introducing another form of bias towards the dataset. Annotators' personal background moderately influenced their decisions and should be taken into consideration when building datasets, along with other factors such as topics, reading news habits, and honest mistakes [28]. Clearly, this is not a cheap procedure and can be a huge bottleneck in creating a reliable media bias dataset.

Alternatively, websites such as Allsides and AdFontes have their own experts and annotations to which can be crawled and make use of (cite cite). However, they operate solely within the U.S., coverting U.S. media sources. To my knowledge, there is no other organisation that do what they do outside the U.S. or in a more global scale, at the same level. Therefore, building a dataset that is not only accurate, but also global and diverse would have massive benefits.

3.2 Classification

Current State-of-the-Art in media bias detection typically to employ neural or transformer-based approach, by fine-tuning or exploiting Large Language Models (cite).

MAGPIE [12] is the first large-scale, RoBERTa-based, multi-task learning (MTL) model dedicated for bias-related task, a promising approach for media bias detection and can be used to enhance the accuracy and efficiency of existing models. Using MAGPIE's context representation instead of BERT for media detection can potentially improve performance. However, currently the model only trained for sentence-level classification and also outputs binary result.

Many past works on automatic detection of media bias (cite cite) also typically operate on a sentence level, complemented by the lack of appropriate and adequate datasets with article-level annotations [6]. They also tend to only output binary result (either biased or not) [32, 18, 17]. One of the main problem with detecting on sentence-level is that multiple sentences within the same article could have opposite or different biases (I think there is a cite for this, I remember reading).

As media content are often delivered in the form of articles containing a number of paragraphs, detection at the article level is far more useful and desirable. More advanced methods could also include global-level spanning over multiple articles by implementing timeline, graphs, or other types of network to capture relationships among articles with same or similar topics.

Using low-level lexical information alone is inadequate for identifying bias within articles, particularly when dealing with a limited dataset [4].

Complex models for long document classification has been shown to not consistently surpass simplistic baseline models in performance [22], although this could be highly-dependant on the dataset used.

Furthermore, The traditional approach of assigning only a single frame label to news articles remains overly simplistic, given that a standard news story often incorporates multiple viewpoints, arguments, or facets, each potentially carrying distinct connotations or framing [31]. An integer label would be a slightly better solution to represent bias from a text, although it is still hardly ideal.

Additionally, it would be good to consider explainability of a model when classifying a bias, as we would need to not just get the result, but to understand why contents are considered bias. Existing media bias detection systems typically concentrate solely on predicting the likelihood of a certain text being biased, offering limited insights into the underlying reason behind the decision.

3.3 Working with article-length text

Article texts length generally falls between medium to long sequences, as they are not as lengthy as legal documents or clinical studies that usually contains pages of text. Most news articles stay between several hundreds to several thousands words (find cite?).

Classification models demonstrate significant results might perform poorly or even fail when they are tested on large documents [33]

The most straightforward approach of standard fine-tuning of an encoder-based langauge model is not necessarily effective as they mostly only able to process a maximum of 512 tokens (BERT), prompting significant information loss. Several techniques have been attempted to address this issue such as the sliding window techniques, CLS techniques [29], where you exploit the 'CLS' token from a BERT embedding, and other more sophisticated models [16].

Additionally, there are other LLMs similar to BERT that are designed to handle longer sequences: Longformer (cite), BigBird (cite). However, most of these models have different architecture than the original BERT and therefore perform differently. Advanced models such as Longformer, ToBERT, and CogLTX do not consistently surpass the baseline models in classification performance, and only performed notably better than the baseline models on only two datasets [22].

Chapter 4

Dataset Building

4.1 BAT dataset

4.1.1 Characteristics

The BAT dataset [27] is chosen instead of NLPCSS [5] for this project due to its article-level suitability and labels fluidity, as well additional metadata as it contains outlets information. It contains 6345 rows of manually labeled news articles from 255 English-speaking news outlets (US-based), originally scraped from Ad Fontes Media's website along with their respective **political bias** and **reliability scores**. Articles in the dataset encompassed a wide arrange of topics such as COVID-19, politics, and lifestyle. The political bias score measures the extent of political influence, ranging from -42 (most extreme left) to +42 (most extreme right). The reliability score reflects the article's truthfulness, with values ranging from 0 (least reliable, containing inaccurate or fabricated information) to 64 (most reliable, original fact reporting). However, since this dataset depends on manual labeling, the selection of articles by Ad Fontes Media likely introduces bias into the dataset [27].

Both political bias and reliability scores on each article were rated using defined metrics and multiple sub-factors, performed by three randomly selected analysts from Ad Fontes Media's team of over 60 experts. The corresponding three scores were then averaged, producing the final article scores. Moreover, each group consists analysts with different beliefs in the political spectrum i.e., left, center, and right.

The reliability score evaluates original fact reporting to analysis, opinion, propaganda, and inaccurate/fabricated information, with scores above 40 generally considered good and scores below 24 typically seen as problematic, scores between 24 and 40 suggest a variety of factors, including a strong presence of opinion and analysis or significant variability in reliability across different articles [8]. This metric is chosen as the main label in this project due to its correlation with textual-level bias: phrasing bias, spin bias, and statement bias described in 2.1

4.1.2 Extension

The original BAT dataset only contains news titles and links (along with other metadata) and is missing the body content of articles. To overcome this, a Python script is written and executed, iteratively visiting each of the URL from the dataset and scrape the news content. This was not an easy task as each website has its own unique structures and formats. Furthermore, the scraped text contains noises that are almost impossible to remove through the script. Some outlets such as The Nation, Chicago Tribune, and Truthout required manual intervention as the scraped text were duplicated over themselves. The current extended dataset contains 5270 rows of articles, mainly due to unavailable websites and missing articles.

To remove noises from article content, the text are then pre-processed extensively. All the content of every article in the dataset were joined into one single list, split into words, and then compared against an English word list (cite), resulting in a list of faulty words sorted by their occurences. Using this list, noisy patterns were analysed and handled through a combination of string and regex methods, conjoined words identified and fixed through a giant Python dictionary. This process is repeated more than several times until contents are valuable enough to work with. Note that at this point some noises still remain within the text as it will take an extensive amount of time and manual labour to completely clean the text.

4.1.3 Analysis

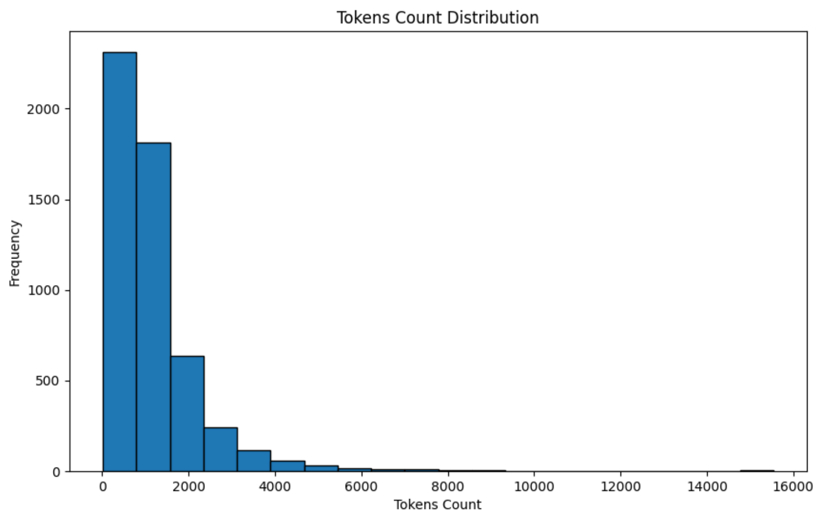

Figure 4.1: Articles tokens count distribution

The article contents tokens length ranges between 22 tokens up to 15530 tokens, with an average length of 1186.5 and a median value of 887 tokens. Only 9 articles have more than 10000 tokens, while there are 106 of articles with less than 100

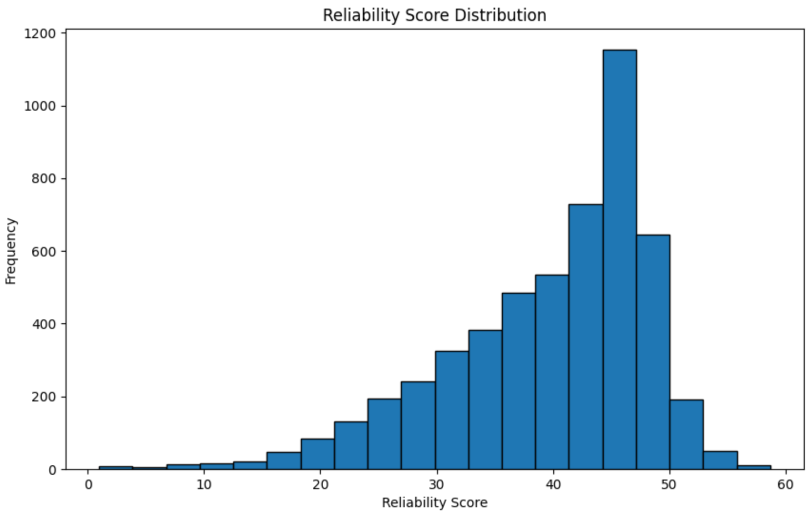

FIGURE 4.2: Reliability score distribution

tokens. Furthermore, only 1193 articles stay between 512 tokens, which is the limit for BERT input. Articles reliability score ranges from 1.0 to the 58.67, the majority have value between 20 - 50. Not a single articles were rated more than 60 despite the highest score being 64. Visualisations can be seen in both Figure 4.1 and Figure 4.2, as well as Figure A.1

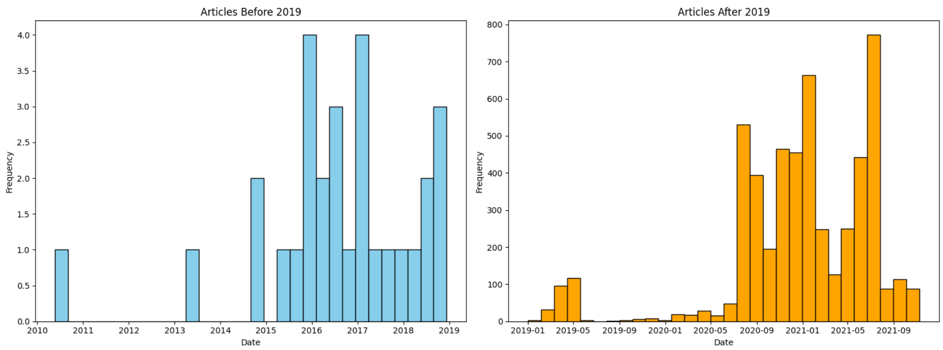

FIGURE 4.3: Article dates distribution

Most articles are written and published within the last 6 years, with only 29 articles, a minuscule percentage, published before 2019, shown in Figure 4.3. From my personal analysis, these 29 articles generally contain similar topics to articles published after 2019 and therefore should not hold any difference in behaviour and characteristics.

Figure A.3 shows that all classes seem to have similar tokens count, close to the overall average. Class 'Problematic' and 'Questionable', being the two most biased

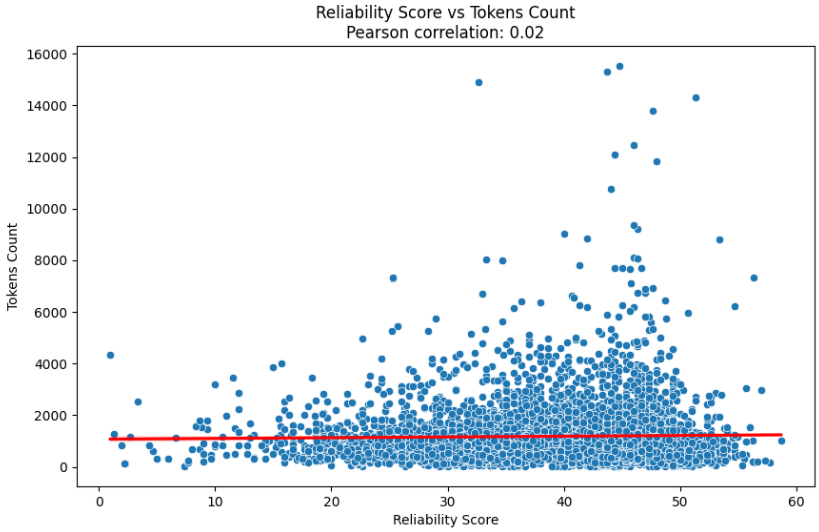

Figure 4.4: Pearson correlation between tokens count and reliability score

classes, seem to have lower average tokens count than the two other classes. However, further analysis (Figure 4.4) shows that there is virtually no linear relationship between between tokens count and reliability score, with a Pearson correlation coefficient is 0.02. This proves that the length of an article has no significant impact with its reliability score. In other words, longer articles are not necessarily more or less reliable than shorter ones based on the provided data.

Chapter 5

Methodology

5.1 Features and baselines

The main focus of the features will be the textual content of articles. A Reliable article-level bias classifier should be able to generalise solely or mainly from the content of the articles, capturing context of the article will be the key element of a reliable performance. Therefore, only articles title and content will be used as the input, concatenated into a single sequence.

As a baseline, traditional methods such as Bag-of-Words and TF-IDF are implemented, combined with a simple logistic regression as a classifier. Standard fine-tuning of BERT will also be evaluated. Additionally, an outlet-based majority votes method is also implemented as a comparison and to show the influence of outlet information within the classifiers. This method works by simply taking the majority class for every outlet and use it as a classifier (if an article is from outlet A then it is class X).

5.2 Pre-processing

The dataset reliability scores are grouped and split into 4 classes based on Ad Fontes split as previously described in Section 4.1.1:

1. Problematic —> scores between 0.00 and 24.00
2. Questionable —> scores between 24.01 and 32.00
3. Generally Reliable —> scores between 32.01 and 40.00
4. Reliable —> scores between 40.01 and 64

The dataset is then split into three sets of train, test and validation with the following distribution:

- Train set: 4145 rows
  270 samples of class 'Problematic', 578 samples of class 'Questionable', 990 samples of class 'Generally Reliable', 2307 of class 'Reliable'

- Test set: 544 rows
  24 samples of class 'Problematic', 51 samples of class 'Questionable', 99 samples of class 'Generally Reliable', 370 of class 'Reliable'

- Validation set: 578 rows
  31 samples of class 'Problematic', 66 samples of class 'Questionable', 123 samples of class 'Generally Reliable', 358 of class 'Reliable'

The split are done in a way to ensure that articles from different outlets are distributed equally between the three sets. This is done by grouping the articles based on its outlet and label and iterating over each group, splitting the rows equally and appending to the train, test, and validation set. Group of less than 5 rows are not enough to be split and therefore appended to the train set.

To handle class imbalances, weighted loss is used when training the model, with weights in proportion the distribution of each class.

5.3 Proposed methods

5.3.1 Sliding window

The first method to be implemented is the sliding window method, where articles are split into chunks and applied as a mini-batch to the model. The logits of each chunks are then pooled together and averaged out to get the final logits, to which the loss function will be applied to. Both max and mean pooling functions were experimented, with mean pooling performing slightly better.

For this approach, a window size of 512 (maximum token of BERT) is chosen with a stride of 256. Additionally, only the first 3 chunks of each input sequence will be used, each input sequence will be truncated to only the first 3 chunks, as using longer chunks do not consistently improve the performance, along with increased computation cost.

5.3.2 CLS method

The CLS method works similarly by splitting each article into chunks. They are first encoded by using a pre-trained LLM (BERT), inputted into the model and taking the last hidden state (as described in [30]) as the text reprensetation. From there, they then get passed into several transformer layers to enhance the contextual representation. Then, for each chunk, only the representation of the CLS token (first token) is then selected, acting as a summary representation of the whole chunk sequence, before finally passing them into a MLP layer. Both LSTM and Bi-LSTM layer were experimented with instead of MLP, with no apparent improvement in performance. MAGPIE and other LLMs can also be used instead of BERT to encode the input sequence.

Chapter 6

Evaluation

All methods are implemented using the PyTorch (cite) and transformer (cite) package from HuggingFace. Batch size is set to 8, with epoch ranging between 4-6 for standard finetuning.

Every method besides the majorty votes only includes textual content as the input sequence (title + content)

6.1 Baseline methods

Class	Precision	Recall	F1	Support
Problematic	0.41	0.29	0.34	24
Questionable	0.35	0.41	0.38	51
Generally Reliable	0.41	0.42	0.42	99
Reliable	0.87	0.85	0.86	370
Overall	0.71	0.70	0.71	

Table 6.1: BoW + logistic regression evaluation

Class	Precision	Recall	F1	Support
Problematic	1.00	0.04	0.08	24
Questionable	0.54	0.27	0.36	51
Generally Reliable	0.43	0.41	0.42	99
Reliable	0.82	0.93	0.87	370
Overall	0.7291	0.7371	0.7070	

Table 6.2: TF-IDF + logistic regression evaluation

In both Table 6.1 and Table 6.2, it can be seen that frequency-based approaches such as BoW and TF-IDF already perform quite well if we look at the average metrics. However, when we look at per class metrics, it can be seen that the overall scores are heavily influenced by the performance on class 'Reliable' due to its large

Class	Precision	Recall	F1	Support
Problematic	0.60	0.25	0.35	24
Questionable	0.47	0.53	0.50	51
Generally Reliable	0.37	0.56	0.45	99
Reliable	0.89	0.79	0.84	370
Overall	0.7428	0.7003	0.7132	

Table 6.3: BERT finetuning evaluation

Class	Precision	Recall	F1	Support
Problematic	0.57	0.71	0.63	24
Questionable	0.60	0.47	0.53	51
Generally Reliable	0.56	0.55	0.55	99
Reliable	0.91	0.93	0.92	370
Overall	0.8007	0.8051	0.8017	

Table 6.4: Outlet-based majority votes evaluation

support. It is quite evident that the model suffers when classifying underrepresented classes, this is expected as the dataset is highly-imbalanced. Furthermore, TF-IDF model seems to perform significantly worse on the ’Problematic’ class.

Finetuning BERT model only performed slightly better than BoW method as can be seen in Table 6.3 and Table 6.1

As a comparison, using solely outlet information without any textual information (with majority votes) outperformed all baseline methods both overall and on every class. This result signifies how influental outlet information can be used to classify media bias.

6.2 Sliding window

6.3 CLS Method

Chapter 7

Conclusion

Appendices

Appendix A

The First Appendix

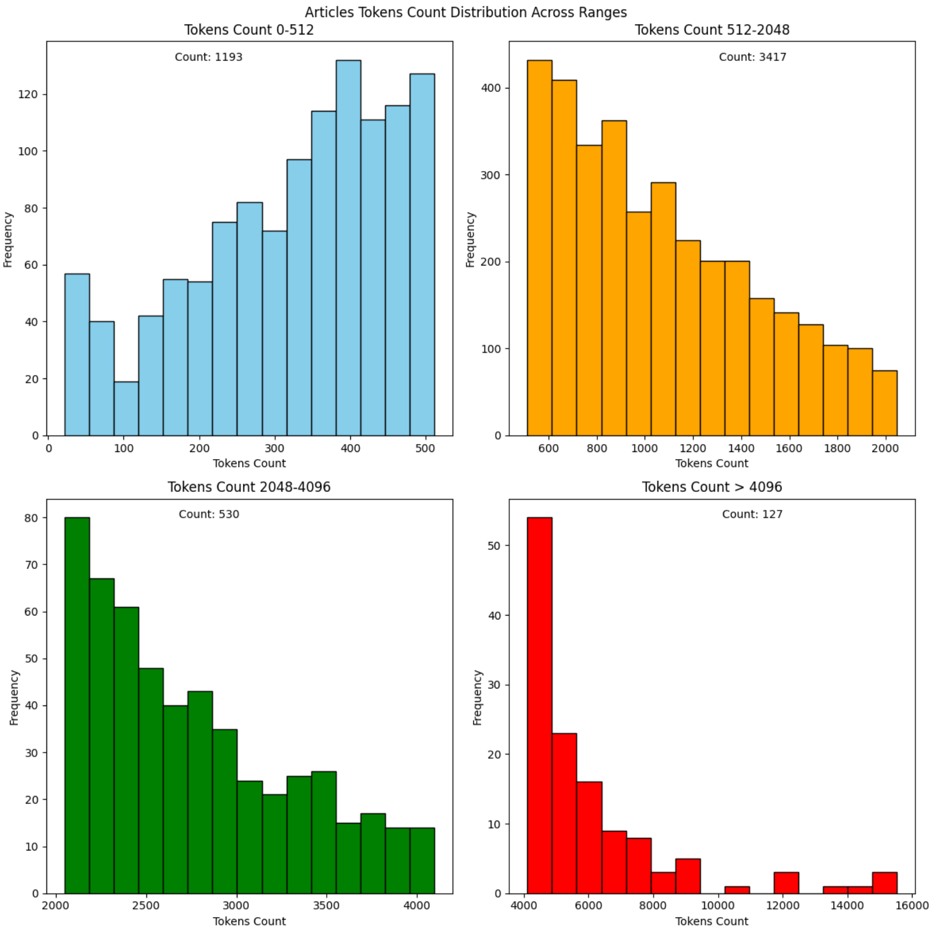

FIGURE A.1: Articles tokens count distribution

Figure A.2: Wordcloud

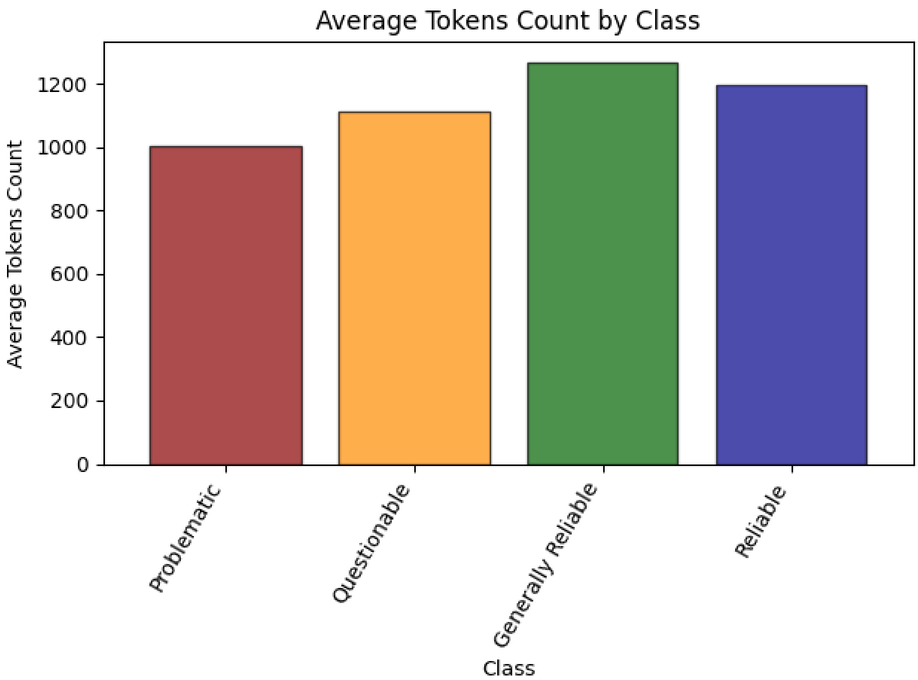

Figure A.3: Average tokens count per class

Appendix B

The Last Appendix

Bibliography

[1] V. P. Aires, J. Freire, F. G. Nakamura, A. S. da Silva, and E. F. Nakamura. An information theory approach to detect media bias in news websites. In *WISDOM 2020*, 2020.

[2] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[3] S. Boudana. A definition of journalistic objectivity as a performance. *Media, Culture & Society*, 33(3):385–398, 2011.

[4] W.-F. Chen, K. Al Khatib, B. Stein, and H. Wachsmuth. Detecting media bias in news articles using Gaussian bias distributions. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online, Nov. 2020. Association for Computational Linguistics.

[5] W.-F. Chen, K. Al Khatib, H. Wachsmuth, and B. Stein. Analyzing political bias and unfairness in news articles at different levels of granularity. In D. Bamman, D. Hovy, D. Jurgens, B. O'Connor, and S. Volkova, editors, *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online, Nov. 2020. Association for Computational Linguistics.

[6] D. Demidov. Political bias of news content: Classification based on individual articles and media, 03 2023.

[7] L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.

[8] A. Fontes. Home - ad fontes media. `https://adfontesmedia.com//`. Accessed: 2024-05-31.

[9] K. Foundation. American views 2020: Trust, media and democracy. Technical report, Gallup and Knight Foundation, 2020.

[10] J. Gottfried and J. Liedke. Partisan divides in media trust widen, driven by a decline among republicans. Technical report, Pew Research Center, 2021. [Online; accessed 29-April-2024].

[11] M. B. Group. Home - media bias group. https://media-bias-research.org/. Accessed: 2024-05-29.

[12] T. Horych, M. Wessel, J. P. Wahle, T. Ruas, J. Waßmuth, A. Greiner-Petter, A. Aizawa, B. Gipp, and T. Spinde. Magpie: Multi-task media-bias analysis generalization for pre-trained identification of expressions, 2024.

[13] S. jeong Lim, A. Jatowt, and M. Yoshikawa. Understanding characteristics of biased sentences in news articles. In *CIKM Workshops*, 2018.

[14] Julie Mastrine. What is media bias? https://www.allsides.com/blog/what-media-bias, 2022. [Online; accessed 28-April-2024].

[15] S. Keeter, N. Hatley, A. Lau, and C. Kennedy. What 2020's election poll errors tell us about the accuracy of issue polling. Technical report, Pew Research Center, 2021. [Online; accessed 30-April-2024].

[16] V. Kulkarni, J. Ye, S. Skiena, and W. Y. Wang. Multi-view models for political ideology detection of news articles. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[17] Y. Lei and R. Huang. Sentence-level media bias analysis with event relation graph, 2024.

[18] I. Maab, E. Marrese-Taylor, and Y. Matsuo. Target-aware contextual political bias detection in news, 2023.

[19] N. Newman, R. Fletcher, K. Eddy, C. T. Robertson, and R. K. Nielsen. Reuters institute digital news report 2023. Technical report, Reuters Institute for the Study of Journalism, 2023. [Online; accessed 30-April-2024].

[20] N. Newman, R. Fletcher, A. Schulz, S. Andı, C. T. Robertson, and R. K. Nielsen. Reuters institute digital news report 2023. Technical report, Reuters Institute for the Study of Journalism, 2021. [Online; accessed 29-April-2024].

[21] C. Panagopoulos. Polls and elections accuracy and bias in the 2020 u.s. general election polls. *Presidential Studies Quarterly*, 51(1):214–227, 2021.

[22] H. H. Park, Y. Vyas, and K. Shah. Efficient classification of long documents using transformers, 2022.

[23] T. E. Patterson and W. Donsbach. News decisions: Journalists as partisan actors. *Political Communication*, 13(4):455–468, 1996.

[24] P. Rafail and J. D. McCarthy. Making the tea party republican: Media bias and framing in newspapers and cable news. *Social Currents*, 5(5):421–437, 2018.

[25] T. Spinde, S. Hinterreiter, F. Haak, T. Ruas, H. Giese, N. Meuschke, and B. Gipp. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias, 2024.

[26] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, and A. Aizawa. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[27] T. Spinde, E. Richter, M. Wessel, J. Kulshrestha, and K. Donnay. What do twitter comments tell about news article bias? assessing the impact of news article bias on its perception on twitter. *Online Social Networks and Media*, 37-38:100264, 2023.

[28] T. Spinde, L. Rudnitckaia, J. Mitrović, F. Hamborg, M. Granitzer, B. Gipp, and K. Donnay. Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, 58(3):102505, 2021.

[29] X. Su, T. Miller, X. Ding, M. Afshar, and D. Dligach. Classifying long clinical documents with pre-trained transformers, 2021.

[30] C. Sun, X. Qiu, Y. Xu, and X. Huang. How to fine-tune bert for text classification?, 2020.

[31] G. Vallejo, T. Baldwin, and L. Frermann. Connecting the dots in news analysis: A cross-disciplinary survey of media bias and framing, 2023.

[32] E. van den Berg and K. Markert. Context in informational bias detection. In D. Scott, N. Bel, and C. Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.

[33] L. Wan, G. Papageorgiou, M. Seddon, and M. Bernardoni. Long-length legal document classification, 2019.