JAIST Repository
https://dspace.jaist.ac.jp/

Title	テキストの単位の違いを考慮したメディアバイアスの検出
Author(s)	李, 東昊
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	author
URL	https://hdl.handle.net/10119/20524
Rights	
Description	Supervisor:白井　清昭, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Media Bias Detection Considering Text Granularity

2410203 LI Donghao

In the contemporary digital era, the landscape of information consumption has undergone a radical transformation, enabling individuals to access an unprecedented volume of different information with remarkably low temporal costs every day. However, this accessibility is inextricably linked to the pervasive risk of media bias, a phenomenon where information is disseminated in a non-objective manner specifically designed to steer the recipient's perceptions, interpretations, and judgments toward a particular direction. Within this context, political bias poses a particularly severe risk, as it has the capacity to obstruct healthy democratic discourse, distort the fundamental processes of public opinion formation, and exert profound, long-term deleterious effects on electoral behavior and civic engagement. To address these concerns, research on automated media bias detection methods utilizing Natural Language Processing (NLP) technology has flourished. Nevertheless, a significant obstacle to the practical application of these technologies remains: the "inconsistency of granularity of text." This phenomenon occurs during the machine learning phase of bias detection models when the granularity of text in the training data significantly differs from that in the test data. In such cases, the discrepancy in volume of information resulting from the inconsistent granularity leads to a marked degradation in the model's discriminatory performance. This study aims to resolve the inconsistency of granularity of text and enhance the overall performance of automated media bias detection. In this research, the detection of media bias is defined as a binary classification problem, where the system must determine whether a given input text constitutes a "Biased" instance containing bias or a "Neutral" instance characterized by objective reporting.

To achieve effective granularity alignment, this research proposes two distinct modules: the "Text Expansion Module" and the "Text Reduction Module." The Text Expansion Module is specifically engineered for scenarios where the primary unit of the granularity of text is a "sentence." The module identifies the original source article to which the target sentence belongs and expands the textual context by concatenating the sentences that immediately precede and follow the target sentence. To ensure that the expansion process only incorporates contextually relevant information and avoids the introduction of irrelevant information, the module calculates the semantic similarity between the original sentence and its neighboring sentences. Textual expansion is only executed if the similarity score exceeds a pre-defined threshold, which is rigorously optimized using validation data. The Text Reduction

Module is applied when the granularity of text is at the "document" level. In this process, the system leverages a Large Language Model (LLM) provided with sophisticated instructions to identify and remove sentences deemed irrelevant to the assessment of media bias or its underlying characteristics. The module generates a condensed summary or "reduced" version that distills the text down to its most critical informational elements. By employing precise prompt engineering, the module transforms lengthy, redundant documents into concise documents that retain high-density information directly related to bias judgment.

In situations where there is an inconsistency of granularity of text between the training and test datasets, these two modules are utilized to achieve precise granularity alignment through two distinct strategies: the "Single-Approach" and the "Double-Approach." The Single-Approach adjusts the length of the test data only. Specifically, when the training data is document-level and the test data is sentence-level, the text expansion module is employed to expand the test data. Conversely, if the training data is sentence-level and the test data is document-level, the text reduction module is used to condense the test data. The Double-Approach, on the other hand, involves the simultaneous adjustment of both the training and test data. By applying the text reduction module to document-level texts and the text expansion module to sentence-level texts across both datasets, the approach seeks to bring both textual sets toward a standardized, median length.

After the adjustment of the text length through these proposed modules, the refined text is fed into an LLM to be converted into an "intermediate representation" that describes the bias with respect to multiple dimensions. These dimensions include tone and language, sources and citations, coverage and balance, agenda and framing, and examples and analogies. Rather than attempting to predict a binary label directly from raw text, this methodology utilizes the intermediate representation to ensure a high degree of explainability regarding the final judgment. The core innovation of this research lies in the dynamic adjustment of the input text's granularity prior to the generation of the intermediate representation. Furthermore, another improvement in this study is the adoption of a configuration where the intermediate representation is input into BERT, acting as a classifier, rather than relying on the simple majority voting method utilized in the baseline model, where samples similar to the test data are retrieved from the training set and the final label is determined by a majority vote.

To evaluate the effectiveness of the proposed methods, four public datasets are used: BABE and BASIL as the sentence-level datasets, while FlipBias and MFC as the document-level datasets. The performance of the models is evaluated using four evaluation metrics: Precision, Recall, and the F1-score,

as well as the Macro F1-score.

The experimental results demonstrated the effectiveness of the proposed Text Expansion and Text Reduction modules in resolving the inconsistency of granularity of text. Specifically, in cases where the training data was at the document-level and the test data was at the sentence-level, it was confirmed that the Double-Approach that adjusted the granularity in both the test and training data significantly improved bias detection performance. Additionally, for cases where the training data was at sentence-level and the test data was at document-level, the results revealed that the Single-Approach that utilized text reduction on the test data worked well for bias detection. These findings underscore the critical importance of granularity alignment of texts in the training and test data.

Furthermore, an ablation study was conducted to verify the individual contributions of each component within the proposed method. Regarding the refinement of the judgment model, it was confirmed that replacing the simple majority voting method used in the baseline with BERT as a classifier significantly contributed to the improvement of detection performance. In addition, the effectiveness of the Text Expansion and Text Reduction modules themselves was validated through the ablation study. Quantitatively, it was shown that excluding these modules prevented the resolution of the inconsistency of granularity of text, subsequently leading to a decline in media bias detection performance.

Finally, a detailed case study and error analysis were performed to clarify the limitations and advantages of the proposed method. While the Text Expansion Module succeeded in supplementing context and background information, it was found to trigger a "dilution effect" in some biased samples, where the additional information tends to lower the relative density of the biased text. Conversely, while the Text Reduction Module effectively stripped away unnecessary information from long documents, it suggested a "neutrality vulnerability" wherein the loss of neutral context increased the likelihood of neutral instances being misclassified as biased. These insights provide valuable guidance for model design in future research.

The major contribution of this thesis is to introduce a new perspective, granularity alignment of text, to the field of media bias detection and to demonstrate its effectiveness from both quantitative and qualitative analysis.