Automating the Creation of Bias Lexica

Harshil Jagadishbhai Darji
University of Passau
Passau, Germany
darji01@ads.uni-passau.de

Shrikanth Singh Balaji Singh
University of Passau
Passau, Germany
balaji01@ads.uni-passau.de

Muhammad Arsal Munir
University of Passau
Passau, Germany
munir01@ads.uni-passau.de

ABSTRACT

In this era of technology, people heavily depend on various media sources to communicate, to broadcast information, and to get updates on what is happening around them, etc. For this reason, media bias can greatly influence the common perception of the topics. People cannot surely identify whether the article they are reading is biased or not until they read a different side of the same story. For the reference sources, such as encyclopedia and scientific texts, the information provided must be neutral. For example, Wikipedia has a core policy, *Neutra Point of View*, which requires editors/contributors to proportionally share all the possible representations of a story, without any bias. As stated by F. Hamborg et al. in [1], researchers in social sciences have developed comprehensive models to effectively describe media bias, but they are often manual. In comparison, models in computer science are fast, automated, and scalable but are simpler.

In our work, we focus on developing a supervised learning approach, that can classify biased statements by leveraging automatically created bias lexicon. This bias lexicon will be generated by training a word embedding method, *word2vec*, on a dataset with chances of having a higher number of biased words. We plan to incorporate this lexicon of biased words with semantic characteristics of a statement to improve the performance of our supervised learning algorithm.

1 INTRODUCTION

Bias usually refers to the inclined point of view towards someone or something. In the case of media bias, this slant can be observed when a media source, such as a news channel favors one person over another or is only interested in reporting the negative side of a story while completely ignoring the encouraging side of it.

For example, consider the following titles of two articles from Daily Mail published online in 2019:

> SARAH VINE: How Kate went from drab to fab! From eyebrows and pilates to a new style guru, our experts reveal the Duchess of Cambridge's secrets to looking sizzling
>
> SARAH VINE: My memo to Meghan Markle following her Vogue editorial - we Brits prefer true royalty to fashion royalty

Both the articles were written by the same editor, published on the same site, but preferred one person over another. In the first article[1], published on 16 June 2019, Kate Middleton was praised for her fashion sense while in the second article[2], published on 30 June 2019, Meghan Markel did not receive the same appreciation over a similar topic.

As explained in [2], [3], and [5], media bias can be explained in different ways:

- *Selection bias*: This happens when a reporter has to decide whether an event should be reported or not.
- *Coverage bias*: This refers to the amount of attention given to a report. It might be the length of the article, reported aspects, or opinions.
- *Framing bias*: Corresponds with the positive, negative, or neutral tone of the statement. For example, the availability of praising or encouraging words in a statement.
- *Epistemological bias*: It focuses on the believability of a statement by checking whether the propositions that are presumed in the text are uncontroversially accepted as true.

In addition to the above-mentioned kinds of bias, we can also observe *statement bias*, *information bias*, and *language bias* in many media articles. Quantifying such bias is harder compared to identifying them in media articles. The main reason is bias in the text is not explicitly given as an opinion or a comment, but rather is subtle and underlying [3].

For this reason, our objective is to automate the process of creating bias lexica, and using this automatically generated bias lexica to classify biased statements using a supervised learning algorithm. In our project, we will use *word2vec* word embedding method to extract biased words from the dataset, then this extracted biased words will be used as seeds to extract more biased words to effectively generate a comprehensive lexicon of biased words.

2 PLAN

Our initial approach[3] to create a lexicon of biased words is outlined in the figure 1.

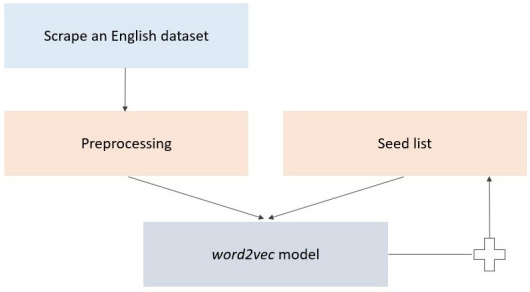

Figure 1: Initial approach to create bias lexica

[1] https://www.dailymail.co.uk/femail/article-7143445/SARAH-VINE-experts-reveal-Duchess-Cambridges-secrets-looking-sizzling.html
[2] https://www.dailymail.co.uk/debate/article-7298911/SARAH-VINE-memo-Meghan-Markle-Brits-prefer-true-royalty-fashion-royalty.html

[3] This will be improved to meet the necessary project needs that may arise in the future and will be explained in more detail as the project progresses.

2.1 Scrape an English dataset

There are different types of datasets we can scrape for our purposes, such as Wikipedia articles, news articles, or blogs.

2.1.1 Wikipedia. Although Wikipedia has an NPOV policy, there still exist tens of thousands of articles flagged with violation of NPOV. We can scape these articles for our project purpose as proposed by C. Hube et al. in [2].

2.1.2 News articles. New articles are a good source of biased words as many of these articles are mostly politically biased. Compared to Wikipedia, these articles are mostly written by professional editors and are regularly updated with new information. For this reason, we are **currently** planning to focus only on news articles.

2.2 Preprocessing

Once the dataset is prepared, the next step will be of preprocessing to remove any unnecessary information such as hyperlinks, emojis (*if any*), or stop words, etc.

2.3 Seed list

This step requires a minimal manual effort. The seed list is a collection of initial biased words that can be used to extract more biased words from the corresponding word vector space. For each word in our seed list, we manually go through the list of closes words in their word representation and pick words that closely represent our seed word. Once we have enough biased words in our seed list, we can use this seed list to automatically extract new biased words from our dataset using *word2vec* word embedding method.

2.4 *word2vec* model

word2vec is a popular word embedding technique capable of capturing the context of a word in surroundings proposed by Tomas at el. [4] in 2013. Word embeddings using *word2vec* can be obtained in two ways: *Skip-Gram* and *Continuous Bag-of-Words* (CBOW).

2.4.1 Skip Gram. As shown in the figure 2, the Skip-gram model predicts words in a certain range before and after the current word.

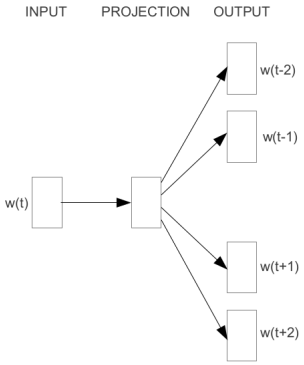

Figure 2: Skip-gram [4]

As stated in the paper [4], increasing the range improves the quality of the resulting word vectors, but increases complexity. It is preferred when the size of the dataset is small.

2.4.2 Continuous Bag-of-Words. In contrast to the Skip-gram model, the CBOW model (figure 3) takes the context of each word as the input and tries to predict the word based on this context.

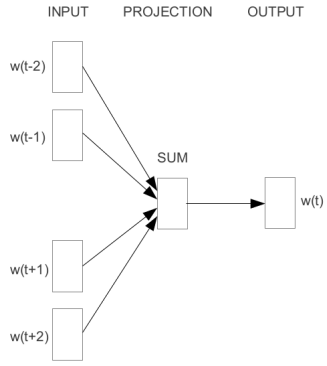

Figure 3: CBOW [4]

CBOW is faster compared to the Skip-gram model and provides better representations for more frequent words.

As stated in *Automating Bias Lexica*[4], in any word2vec model, words frequently appearing in a similar context may lead to words that are not biased. To overcome this issue, the mean of 10 randomly chosen words from the seed list can be used to compute the most similar words for a batch of words.

As mentioned before, once this lexicon is complete, we will try to classify biased statements by incorporating this newly created bias lexicon with semantic characteristics of the statement. This method has proven to achieve 73% accuracy on the Wikipedia dataset [2].

REFERENCES

[1] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2018. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20 (Nov. 2018), 391–415. https://doi.org/10.1007/s00799-018-0261-y

[2] Christoph Hube and Besnik Fetahu. 2018. Detecting Biased Statements in Wikipedia. In *WWW '18: Companion Proceedings of the The Web Conference 2018.* 1779–1786. https://doi.org/10.1145/3184558.3191640

[3] K. Lazaridou, R. Krestel, and F. Naumann. 2017. Identifying Media Bias by Analyzing Reported Speech. In *2017 IEEE International Conference on Data Mining (ICDM).* IEEE, New Orleans, LA, 943–948.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints* (Jan. 2013), 391–415. arXiv:1301.3781

[5] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. *ACL 2013 - 51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1650–1659.

[4] https://bit.ly/2nwHiFU