Towards Bias Detection in Online Text Corpora

Christoph Hube[1], Robert Jäschke[2] and Besnik Fetahu[1]

[1] L3S Research Center, Leibniz Universität Hannover, Germany
{hube, fetahu}@L3S.de
[2] Institut für Bibliotheks- und Informationswissenschaft, Humboldt University Berlin, Germany
robert.jaeschke@hu-berlin.de

Abstract. Natural language textual corpora depending on their *genre*, often contain bias which reflect the point of view towards a subject of the original content creator. Even for sources like Wikipedia, a collaboratively created encyclopedia, which follows a *Neutral Point of View* (NPOV) policy, the pages therein are prone to such violations, this due to either: (i) Wikipedia contributors not being aware of NPOV policies or (ii) intentional push towards specific points of views. We present an approach for identifying bias words in online textual corpora using semantic relations of word vectors created through word2Vec. The bias word lists created by our approach help on identifying biased language in online texts.

1 Introduction

To enforce neutrality and quality of the provided information, Wikipedia has established several guidelines and policies. *Neutral Point of View* policy demands Wikipedia editors to put aside their personal opinions on a topic and create objective content. Even for information sources that allow opinions or where opinions are part of the sources' agenda (e.g. many news websites) it is helpful for the readers to understand the intrinsic bias of sources. Especially in the context of *filter bubbles* and *echo chambers* [1,7], bias detection plays an important role. In this work we aim to detect automatically the use of explicit language bias, i.e. bias that is introduced through specific words and phrases. Language bias stands in contrast to more implicit bias that is introduced through *gatekeeping* or coverage of specific topics [5]. As an example of language bias consider the following two sentences:

- *"Barack Obama served as president of the United States."*
- *"Barack Obama* ***unsuccessfully*** *served as president of the United States."*

The first sentence follows a neutral point of view. In the second sentence bias is introduced by adding the word *unsuccessfully*.

We call words that typically introduce bias to a statement or are a strong indicator of bias in a statement *bias words*. Bias words can be grammatically diverse with existing examples across nouns, verbs, adjectives, adverbs and more, and furthermore they may vary based on the context they occur. In this paper we present an approach for identifying bias words in online text corpora using the semantic relations of word vectors created through word embedding approaches like *word2Vec* [2]. The resulting bias words can be used for bias detection in text.

Recasens et al. [4] tackle the language bias problem where they identify the most biased word in a sentence already knowing that the sentence is biased. To do so, they rely on language features such as lists of factive, assertive and implicative verbs, and additionally make use of a bias lexicon extracted from a subset of Wikipedia revisions, in which the editor mentions the abbreviation POV (Point of View). In contrast to Recasens et al. [4], our approach differs in that we provide a comprehensive list of bias words with nearly $\sim$ 10,000 words, and in that we make use of word embeddings, which capture semantics and syntactic relationships between words, to extract words that may indicate bias.

In Section 2 we introduce a semi-automated approach for seed word extraction (Section 2.1) and a fully automatic approach for extracting bias words given a set of seed words and a fitting text corpus, e.g. the latest Wikipedia corpus (Section 2.2).

2 Approach

In this section, we describe in detail the two main steps of our approach: (1) seed word extraction, and (2) bias word extraction.

2.1 Seed Word Extraction

Through empirical observations, we see that bias words often co-occur with other bias words in the word vector space. In order to identify these bias word clusters, we first need to extract a small number of bias words that we can use as seeds for our approach. The idea is to use word vectors that already have a high density of bias words since it will make the manual identification of bias words faster. Therefore we use a corpus from which we expect to have a high density of bias words compared to Wikipedia.

Conservapedia[1] is a Wiki shaped according to right-conservative ideas including strong criticism and attacks especially on liberal politics and members of the Democratic Party of the United States. Since no public datset is available we crawled all Conservapedia articles under the category *politics* (and all subcategories). The crawled dataset comprises of a total of 11,793 articles. We preprocess the data using a Wiki Markup Cleaner. We also replace all numbers with their respective written out words, remove all punctuation and replace capital letters with small letters. In the next step we use word2Vec to create word embeddings based on the Conservapedia dataset.

To achieve a high density of bias words, we explicitly pick words that are associated with a strong political split between left and right in the US (e.g.

[1] `http://www.conservapedia.com`

media, immigrants, abortion) for the seed word extraction. We leave for future work to automate the process of seed word extraction, where approaches like [3] can serve as a starting point, however, its use can be limited since clean and explicit labels (bi-partisan or more POVs) of the textual corpora is required, a task deemed to be very difficult considering the broad coverage in encyclopedias like Wikipedia.

For each word we then manually go through the list of closest words in the vector space using cosine similarity and extract words that seem to convey a strong opinion. For example among the 100 closest words for the word *media* in the vector space we find words such as *arrogance*, *whining*, *despises* and *blatant*. We merge all extracted words into one list. The final seed list contains 100 bias words.

2.2 Bias Word Extraction

Given the list of seed words, we extract a larger number of bias words using the Wikipedia dataset of latest articles[2]. We preprocess the dataset in the same way as we preprocessed the self-crawled Conservapedia dataset and create a word vector space using word2Vec. First, we split the seed word list randomly into $n = 10$ batches of equal size. In the next step we use the semantic relations of word vectors created to identify clusters of bias words. For each batch of seed words we sum up the word vectors of each word in the batch. Next, we extract the closest words according to the cosine similarity of the combined vector. By using the combined vector of multiple seed words we increase the probability of extracting bias words compared to the use of only one seed word. Table 1 shows an example of the top 20 closest words for the single seed word *indoctrinate* and a batch containing *indoctrinate* and 9 other seed words. Our observations suggest that the use of batches of seed words leads to bias word lists of higher quality.

We use the extracted bias words as new seed words to extract more bias words using the same procedure. Table 2 shows statistics for our extracted list of bias words. The list contains 9742 words with 42% of them tagged as nouns, 24% tagged as verbs, 22% tagged as adjectives and 10% tagged as adverbs. The high number of nouns is not surprising since nouns are the most common part of speech in the English language. To annotate the words with their part of speech we use the POS tagger[6]. We provide the final bias word list at the paper URL[3].

3 Conclusion and Future Work

We introduced a new approach for extracting bias words using word2Vec from textual corpora like Wikipedia. We are planning to integrate bias word lists among other features into a machine learning classifier for bias detection. For a

[2] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

[3] https://git.l3s.uni-hannover.de/hube/Bias_Word_Lists

Table 1. Top 20 closest words for the single seed word *indoctrinate* and the batch containing the seed words: *indoctrinate, resentment, defying, irreligious, renounce, slurs, ridiculing, disgust, annoyance, misguided*

Rank	Single seed word	Batch of seed words
1	cajole	hypocritical
2	emigrates	indifference
3	ingratiate	ardently
4	endear	professing
5	abscond	homophobic
6	americanize	mocking
7	reenlist	complacent
8	overawe	recant
9	disobey	hatred
10	reconnoiter	vilify
11	outmaneuver	scorn
12	helmswoman	downplaying
13	outflank	discrediting
14	renditioned	demeaning
15	redeploy	prejudices
16	seregil	humiliate
17	unnerve	determinedly
18	titzikan	frustration
19	unbeknown	ridicule
20	terrorise	disrespect

Table 2. Statistics about the extracted bias word list

POS tag	#	ratio
nouns	4101	(42%)
verbs	2376	(24%)
adjectives	2172	(22%)
adverbs	997	(10%)
others	96	(1%)
total	9742	

proper evaluation we will use crowdsourcing to generate a ground truth of biased and non-biased statements from both Wikipedia and Conservapedia.

Acknowledgments This work is funded by the ERC Advanced Grant ALEXANDRIA (grant no. 339233), DESIR (grant no. 31081), and H2020 AFEL project (grant no. 687916).

References

1. R. K. Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, 14(2), 2009.
2. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
3. B. L. Monroe, M. P. Colaresi, and K. M. Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.
4. M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In *ACL (1)*, pages 1650–1659, 2013.
5. D. Saez-Trumper, C. Castillo, and M. Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In *22nd CIKM*. ACM, 2013.
6. K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 63–70, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
7. V. Vydiswaran, C. Zhai, D. Roth, and P. Pirolli. Biastrust: Teaching biased users about controversial topics. In *21st CIKM*. ACM, 2012.