Bias Detection and Neutralization of Wikipedia Articles

Harsha Bang[1], Piyush Muthal[2], Snehal Brahmane[3], Komal Kamble[4]

[1]*Graduate Student, Dept. of Computer Engineering, MES College of Engineering Pune – 01, India*
[2]*Graduate Student, Dept. of Computer Engineering, MES College of Engineering Pune – 01, India*
[3]*Graduate Student, Dept. of Computer Engineering, MES College of Engineering Pune – 01, India*
[4]*Graduate Student, Dept. of Computer Engineering, MES College of Engineering Pune – 01, India*

Abstract—*Biased means the perceptual strong instinct or a judgment on something. Bias articles are introduced by unique offensive words or phrases during a declaration and must be excluded in order to form the objective declaration. Wikipedia focuses on the policy of the Rational Point of View that implies that Wikipedia data must be neutral. Due to sizable amount of articles on Wikipedia and operating guidelines with voluntary basis of Wikipedia editors, the quality assurance and Wikipedia guidelines can not be always followed. There has been many researchers to developed different techniques ideas to get rid of subjectivity during this paper we are that specialize in those techniques and briefly study about them.*

Keywords—Neutral Network, Random forest, Naive Bayes, Logistic Regression, Support Vector Machine, NLP.

1. INTRODUCTION

Since Wikipedia is primarily created by users, it is assumed that the expression of viewpoint is expected. Wikipedia follows a Rational Point of View policy according to which documents should, to the extent possible, be objective, appropriate and bias-free. Wikipedia's policy document advises editors to avoid presenting as an opinion unproblematic facts and, on the other hand, to avoid stating personal views or disputed statements as facts, historical narrative, from the very first field study to the present day[1].

Understanding objective and subjective terminology and the distinction between the two allows us to make informed decisions on data on the concept of subjectivity and the need to neutralize essential and important documents. Subjective is a perception of something. Whereas Objective refers to something which is not influenced nor interpreted by others opinion. This prejudice is introduced by provocative terms and phrases in natural language, putting doubt on evidence and assuming the truth.

[2].Every language plays a very significant role to balance our communication and represent our point of view by expressing our thoughts, sharing ideas with others. It mostly depends on our personal experiences and our own perspective. Unbiased language is an important part of balanced and fair representation in writing. Biased language can discriminate between the opinions, demean or offend the society which won't be acceptable in news, Wikipedia articles according to the guidelines. We should make sure that the review or the content does not demean or offend anybody's thoughts or opinions. Usually, it is influenced by emotions or opinions. Statements and terminology in collaborative contexts or environments where objective language is required (e.g. Wikipedia, news media) should be equally interpreted by the parties concerned and neutrally articulated. Through the presence of offensive terms or phrases, or comments that may be wrong or one-sided, biased language is adopted, thereby breaching such consensus[3].The vocabulary used in Wikipedia should be impartial [4]. The processing of natural language sets such as word embedding increases the accuracy of predictive models. The word embeddings often include and amplify biases such as stereotypes and prejudice[5] present in data.

1.1 TYPES OF SUBJECTIVE BIAS

As shown in Fig1.there are 3 types of subjective sentences:

Epistemological Bias: Linguistic features that concentrate subtly on the credibility of the proposal. Example: I Piyush Goel said that all Indian villages were electrified and tended to favor liberal viewpoints. This can be written as 'Piyush Goel said that all Indian villages were electrified and tended to favor liberal viewpoints[6].

Demographic Bias: Presumptions about highly relevant gender or other demographic categories. Example: i) In general, an IT consultant spends his career mired in darkness. It can be corrected as '**IT consultants** often spend their careers mired in darkness.'[6]

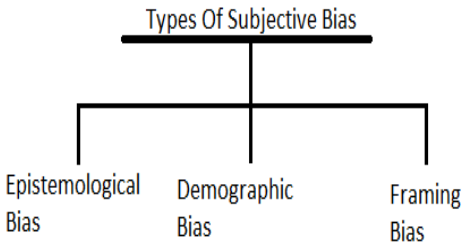

Fig1.

Framing Bias: Framing bias happens when individuals make those choices based on the way evidence is framed, as opposed to only the facts themselves. That is the same fact presented with different ways can lead to different judgements. In simple words it is the interpretation of facts by one's subjective opinion.[1]

Other forms of bias, such as selection bias and bias based on particular subjects, such as gender bias or cultural bias, are also addressed in some studies. Open views were not found biased by Hube C and Fetahu B. For instance, the argument that I think this movie is very bad is not, by definition, biased because the author makes it clear that it is her own opinion.

Even though there are more types of subjectivity bias, the categories in which the subjectivity can be included, need to be studied and understood. These categories can be separated or identified by analyzing the pattern of editing the articles. This can be done by using WNC corpus. As Wikipedia follows the NPOV corpus, Reid Pryzant Alt studied the WNC where they ignore some edits to maximize the precision, like [1].

- The edits where more sentences were changed.
- The one in which nouns are more than half of the words in the sentence.
- The edits in which spelling or grammatical errors are present.
- Edits that includes references or hyperlinks.
- The one in which the paragraph has symbolic elements, such as tables or punctuation

1.2 Wiki Neutrality Corpus (WNC) :

Pre and post-neutralization are part of the Wiki Neutrality Corpus.It consists paragraphs by English Wikipedia editors[1]. The corpus has been extracted from Wikipedia edits that are designed to ensure a rational point of view for writings. WNC is the first parallel corpus that targeted on biased and neutralized language. In WNC,the type of subjective bias in sentences is determined.It helps to understand the characteristics of the subjective bias on Wikipedia.Subjectively biased editions are more popular in history, politics, philosophy, athletics, and language categories. They are less common in the categories of meteorology, science, landforms, broadcasting and arts[1]. This suggests that there is a connection between the topic of a text and the exposure of bias. We can work on categorical bias neutralization by using the data provided in WNC. Since the NPOV corpus version is primarily designed to eliminate the subjectivity with the help of logistic regression and linguistic features including factive verbs, hedges and subjective intensifiers which are used to detect bias-inducing words[2]. They have used some methods that include multi-word edits which detect sentence-level bias.

1.3 Word Embedding:

A word embedding is a learned text representation in which a similar representation is provided to words that have the same meaning. This approach to the representation of words and documents can be considered one of the main advancements of deep learning on difficult issues related to the processing of natural language.Word embedding is a set of techniques of natural language processing that explain words to real vector numbers. These vectors are used to raise the standards of relational and qualitative models[5].The use of word embeddings to form a model which is highly precise to perform text generation, translation, classification and regression without taking into account the effect of their inherent biases [5].BERT (Bidirectional Encoder Representations from Transformers) models were pre-trained with big dataset of sentences. Important work has been done to detect subjectivity using text classification models ranging from linguistic features based models focused on characteristics to fine tuned pre-trained word embeddings such as BERT[4]. It is a powerful tool for NLP for calculating similarity measures within the sentences.

1.4 Linguistic Features:

The linguistic characteristics of the various form of bias categories differ according to the context of the article. These characteristics can be captured on the idea of the cues encountered during the training of datasets. The context must be analyzed, as biases can be rely on the context, especially epistemological bias, as they depend on the fact of the proposal[6]. There is a sociolinguistic theory in which language and linguistic structure are the medium for the function of a specific social group. Language represent the group's parameters and other characteristics (i.e., ideology, economical, cultural). This generally leads to public consensus on the usage of terminology on a specific topic and the meaning of particular phrases and words[3].In their work, Rohit Raj and Rahul Agarwal listed the kinds of features used in the

logistic regression model along with their significance domain. The count for it is 36,787.It includes previous revisions of the documents and the ratio between the number of times the word was altered by the neutralization term and its incidence frequency. The aim of this feature is to remove framing bias. For linguistic theory, understanding linguistic bias is essential; equally critical for computational linguistics is the computerized detection of biases. The edits related to the NPOV tags allow us to identify the text in its biased (before) and objective (after) form, letting us understand the linguistic realization, as Wikipedia keeps the overall history of revised words.

2. RELATED WORK

Christoph Hube and Besnik Fetahu explained [3] the RNN based approach was used for classifying statements that contained biased language. They focused on the case of biased phrasing, that were the statements in which words and phrases were inflammatory or partial. The representation of words in a phrase was an important prerequisite for the effective implementation of RNN models in their assignment. The three main phrase representations have been differentiated. They had distinguished three main sentence representations that were Word Representation, POS Tags and LIWC Word Functions. In contrast with feature-based models, the RNN models were superior in output and were able to catch the significant terms and phrases that introduced bias in their statement. With a very high accuracy of 91.7 percent, they were able to predict the bias.

Reid Pryzant Et al. [1] for this neutralization mission, they had proposed a pair of sequence-to-sequence algorithms. Both strategies exploit autoencoders and token-weighted loss function denoising. The algorithm had splitted their problem into (1) identification (2) editing, specifically identifying troublesome words using BERT-based detector and a novel join embedding in which the detector could modify an editor's hidden states. This paradigm encouraged an important human-in-the-loop approach to understand the bias and modeling generative language. Secondly, it was easy to train and use, but the "CONCURRENT" method was not transparent. BERT encoder was used as part of the generation process to define subjectivity. Also they used LSTM-based editing. The pretrained model from each stage was combined into a one system. But the scope was limited to single-word edits, which focused only on the quarter of the edits in their data, and was probably for the simple bias sentences[1].

Oriestis Papakyriakopoulos Et al. [5] explained a new technique[5] for gender language bias detection and which was used to analyze biases in Wikipedia-trained embeddings and political social media data. The results were divided into three sections in this paper, firstly, they presented findings on Wikipedia bias and word embeddings for social media..Secondly, they studied how the biasness was distributed and how to minimize it.When used in sexism detection models, they also demonstrated the efficiency of biased word embedding. The assessed bias in word embedding was further diffused in the changed sentiment classifiers in the last. Each embedding had one classifier trained, with accuracy of around 85%.The drawbacks were not found there because the semantic quality of words has always been related to a society's sociopolitical ties and reliance on the existence of the input data on word embedding[5].

Desislava Aleksandrova Et. al.[8] proposed a multilingual method for extracting Wikipedia sentences and used them to create corpus in Bulgarian, French and English.. The hypothesis was that having similar examples in both bias and unbiased classes would help to identify discriminatory words targeted by NPOV-related edits. As this method did not rely on language-specific features other than the NPOV tag list and a stop word list, it was easily applied to Wikipedia archives in other languages[8].

Marta Recasens Et al.explained [7] Actual bias and bias-driven edits extracted from Wikipedia. For each word that initially appeared in the NPOV sentences of the training set, was trained on a logistic regression model , with biased words as a positive class, and all the other words as a negative class. At the time of the test, a set of sentences was given to the model and, for each of them,the words were placed as per their probability of being biased.

Rohit Raj and Rahul Agarwal [6] explained the importance of detecting biases in various articles .Since unbiased language was very important to be followed for sources such as news articles, Wikipedia articles. As for the Wikipedia policy of neutral point of view i.e. (NPOV) that suggested that articles should be declared impartial. This paper helped to detect the biased sentences in the articles.They have written in python3, all the scratch modules that were trained and many models such as linear regression, SVM, CRF using the sklearn library. It was modelled as a sequence labeling problem where each word was categorized in O(Unbiased) or B(Biased) in a phrase, then they used a CRF for the labeling task[6].

Tanvi Dadu Et. Al. [2]the implementation of BERT-based models to the task of subjective language detection was being explained. Numerous BERT-based models have been studied, including BERT, Ro BERT, AL BERT, along with their native classifiers with their base and broad specifications.. They have also provided an ensemble model that used multiple ensemble techniques to give predictions. Their proposed model exceeded the baselines by 5.6% of the F1 score and 5.95% of the Accuracy. FastText, BiLSTM, BERT were its baseline models used in the projects. They also integrated multiword edits by detecting bias at the sentence level[2].

S.No	Paper Title	Drawback
1	Automatically Neutralizing Subjective Bias in Text[1].	This method was limited to single-word edits, where there was just a quarter of the edits in the data and which were among the easiest cases of bias.
2	Towards Detection of Subjective Bias using Contextualized Word Embeddings[2].	This paper has explored only sentence level bias detection using BERT.
3	Neural based statement classification for biased language[3].	As this model has not analyzed the different forms of biases we can not use it for the sentences or paragraphs containing selection and demographic bias.
4	Detecting biased statements in Wikipedia[4].	They could not improve the classification results based on their work and datasets.In addition they can clarify the granularity of classifiers to detect the biases of different languages
5	Bias in Word Embeddings[5].	Word embeddings in their results created algorithmic social discrimination ,which gave particular social classes and person negative inferences.
6	Bias Detection[6].	The SVM worked well on the training data, but not on the test data, the reason was over-fitting. Their model lacked proper dataset due to which accuracy was not up to the mark.
7	Linguistic Models for Analyzing and Detecting Biased Language[7].	The bias lexicon did not helped much as it was over fitted to the training data.
8	Multilingual Sentence-Level Bias Detection in Wikipedia[8].	Due to lack of proper dataset and quality they could not cover issues during human evaluation such as incoherent sentences,segmentation and quality of the dataset.

Hube C. and Fetahu B [4] explained that Wikipedia had a set of editing guidelines and policies for the demographic groups and interests of editors and to achieve the quality of the information provided [4]. In the paper, Wikipedia statements, they addressed had quality problems that dealt with language bias that were in violation of points (i) avoid stating opinions as facts and(ii)prefer non judgemental language. They have structured lexicon of words using word representation techniques such as word2wec which was found to be effective in disclosing words for a particular word that were similar to or used in a similar context. Using two steps(i)Seed word extraction where high-density bias words were extracted from the list(ii)Bias Word Extraction, which extracted words from the list of seed words from which word embeddings were computed using word2Vec and skip-gram model.

3. CONCLUSION

This paper gives an overview about Bias detection in Wikipedia articles. The model discussed in these papers were developed to avoid opinionated thoughts and save editors time . Many points relevant to Wikipedia posts are well known after much study on identifying bias words. Articles such as Wikipedia should be portrayed equally, proportionately, and without any prejudice to the extent possible.

One of our goals in this review was to consolidate existing quantitative results and to carry out comparative analysis. We have successfully studied previous research papers. We have found that the dataset they have been through has a lack of quality and variability and this directly affects the accuracy of the model. Some papers detected the biased sentences without giving any idea of neutralization. While others neutralized them by replacing the bias words. Some papers didn't give the expected results as they lack a proper dataset on which they can test and train their data.

4. REFERENCES

[1]Reid Pryzant, Richard Diehl Martinez, Nathan Dass,Sadao Kurohashi,Dan Jurafsky,Diyi Yang,"Automatically Neutralizing Subjective Bias in Text ",34th AAAI Conference on Artificial Intelligence 2020.

[2]Tanvi Dadu(NSIT Delhi) , Kartikey Pant,Radhika Mamidi "Towards Detection of Subjective Bias using Contextualized Word Embeddings",Web Conference 2020 , Association for Computing Machinery, New York, NY, USA,February 2020.

[3]Hube, C.and Fetahu B. "Neural based statement classification for biased language", In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 195–203. ACM.2019.

[4]Hube C. and Fetahu B. "Detecting biased statements in Wikipedia" ,In The Web Conference, 1779–1786. International World Wide Web Conferences Steering Committee 2018.

[5]Orestis Papakyriakopoulos ,Simon Hegelich ,Juan Carlos, "Bias in Word Embeddings ", In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 2020.

[6]Rohit Raj and Rahul Agarwal, "Bias Detection" CSE Publications IIT Delhi 2018.

[7]Marta Recasens ,Cristian Danescu-Niculescu-Mizil and Dan Jurafsky, "Linguistic Models for Analyzing and Detecting Biased Language". Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 2013.

[8]Desislava Aleksandrova, François Lareau and Pierre-Andre Menard, "Multilingual Sentence-Level Bias Detection in Wikipedia" Conference: Recent Advances in Natural Language Processing At: Varna, Bulgaria 2019.