Multilingual Sentence-Level Bias Detection in Wikipedia

Desislava Aleksandrova [1,2] **François Lareau** [1] **Pierre-André Ménard** [2]
[1] OLST, Université de Montréal, first.lastname@umontreal.ca
[2] Computer Research Institute of Montreal, first.lastname@crim.ca

Abstract

We propose a multilingual method for the extraction of biased sentences from Wikipedia, and use it to create corpora in Bulgarian, French and English. Sifting through the revision history of the articles that at some point had been considered biased and later corrected, we retrieve the last tagged and the first untagged revisions as the before/after snapshots of what was deemed a violation of Wikipedia's neutral point of view policy. We extract the sentences that were removed or rewritten in that edit. The approach yields sufficient data even in the case of relatively small Wikipedias, such as the Bulgarian one, where 62k articles produced 5k biased sentences. We evaluate our method by manually annotating 520 sentences for Bulgarian and French, and 744 for English. We assess the level of noise and analyze its sources. Finally, we exploit the data with well-known classification methods to detect biased sentences. Code and datasets are hosted at https://github.com/crim-ca/wiki-bias.

1 Introduction

Our goal is to automatically detect neutral point of view (NPOV) violations at the sentence level with a procedure replicable in multiple languages. Sentence-level bias detection is a type of sentiment analysis, closely related to subjectivity detection (Riloff and Wiebe, 2003; Wiebe and Riloff, 2005; Wilson and Raaijmakers, 2008; Murray and Carenini, 2009; Lin et al., 2011; Al Khatib et al., 2012), where an opinion is considered subjective, and a fact, objective. Yet, as far as bias in writing is concerned, both subjective opinions and objective fact reporting (cf. §5) may, in some cases, be sources of impartiality. The importance of the context is one of the main difficulties in detecting bias at the sentence level. Some types of point-of-view bias are equally challenging for humans to detect. Partisanship in editorials, for example, tends to go unnoticed when in line with the reader's own ideas and beliefs (Yano et al., 2010). A further complication arises from the ambiguity of the term bias, which stands for a lack of fairness or neutrality in realms as varied as human cognition (Tversky and Kahneman, 1974), society (Ross et al., 1977), media (Entman, 2007), internet (Baeza-Yates, 2018; Pitoura et al., 2018) or statistical models and algorithms (O'Neil, 2016; Shadowen, 2019), to name a few. With so many different types of bias and their varying definitions, it is not trivial to set the scope of a bias-detection study.

The majority of the work on this task is performed on news articles (Hirning et al., 2017; Baly et al., 2018; Bellows, 2018) and political blogs (Yano et al., 2010; Iyyer et al., 2014) rather than Wikipedia, because of the relative scarcity of examples an encyclopedia provides. Yet, unlike alternative data sources, Wikipedia comes with a definition of bias outlined in its content policy for neutrality of point of view (NPOV). The core guidelines in NPOV are to: (1) avoid stating opinions as facts, (2) avoid stating seriously contested assertions as facts, (3) avoid stating facts as opinions, (4) prefer nonjudgemental language, and (5) indicate the relative prominence of opposing views. In addition, Wikipedia provides lists of bias-inducing words to avoid,[1] such as positively loaded language (puffery) in the form of *peacock* words (e.g., *best, great, iconic*); unsupported attributions, or *weasel words* (e.g., *some people say,*

[1] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch

Proceedings of Recent Advances in Natural Language Processing, pages 42–51,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_006

it is believed, science says); uncertainty markers, known as *hedges* (e.g., *very, much, a bit, often, approximately*), editorializing (e.g., *without a doubt, arguably, however*) and more. When an article is considered biased, an editor can flag it by adding a tag such as {{POV}} to its source, which displays a disputed neutrality warning banner on the page. These explicit guidelines (and the editors who apply them) help reduce biased language in Wikipedia over time through a continuous process of collaborative content revision (Pavalanathan et al., 2018). Still, new instances of bias are introduced just as often as old ones are overlooked because of humans' inherent difficulty with subtle expressions of point-of-view partiality. Recasens et al. (2013) showed that when presented with a biased sentence from Wikipedia, annotators manage to correctly identify the loaded word in only 37% of the cases.

2 Related Work

Bias detection approaches vary primarily in terms of corpora, vectorization methods, and classification algorithms. We present a review of the related literature along this division.

2.1 Corpora

Among those who tackle NPOV violations in Wikipedia, some rely on available datasets (Vincze, 2013), others perform manual annotation (Hube and Fetahu, 2018; Ganter and Strube, 2009; Herzig et al., 2011; Al Khatib et al., 2012), still others attempt to automatically extract labeled examples (Ganter and Strube, 2009; Recasens et al., 2013; Hube and Fetahu, 2018). Our approach is in line with the latter.

Using existing corpora, while being the cheapest method, predetermines which types of bias will be explored and in which languages. Vincze (2013) uses WikiWeasel, the Wikipedia subset of the CoNLL-2010 Shared Task corpora (Farkas et al., 2010) to study discourse-level uncertainty by manually annotating linguistic cues for three overt manifestations of bias: weasel, hedge and peacock words. Ganter and Strube (2009) focus on detecting hedges in a corpus of 1000 extracted sentences tagged with {{weasel}}, Bhosale et al. (2013) try to detect promotional content, while Kuang and Davison (2016) train their model on the English corpus of Recasens et al. (2013).

Manual annotation ensures higher quality but is too costly for large multilingual datasets. Hube and Fetahu (2018) learn to detect bias in Wikipedia on a manually annotated corpus of sentences from the inherently biased Conservapedia, with a precision of 0.74. When tested on an unlabeled dataset extracted from Wikipedia however, the classifier obtains a precision of 0.66 for the sentences classified with a certainty over 0.8.

Recasens et al. (2013) first propose a heuristic to automatically build a labeled corpus with biased sentences. Out of all revisions of NPOV-tagged articles, they identify the bias-driven edits based on the comments the editors left at commit. Although reliable, this method yields a fairly small set of examples for English (2,235 sentences) and none for smaller Wikipedias, first because of its dependence on revision comments (which are optional), and second, because it limits the examples to bias-driven edits containing five or fewer words.

2.2 Vectorization

As for data vectorization, previous work on bias detection relies either on features from pre-trained language models, custom feature-engineering or both. Bellows (2018) finds no significant difference in performance for classifiers trained on Word2vec, GloVe, or fastText representations. Several studies (Recasens et al., 2013; Ganter and Strube, 2009; Hube and Fetahu, 2018) employ multiple lexical, contextual, and linguistic features which, while boosting performance, remain dependant on handcrafted word lists, specialized lexical resources such as SentiWordNet (Baccianella et al., 2010), subjClue (Gitari et al., 2015), etc., and grammatical parsers that often cover only English. Yano et al. (2010) combine word vector representations from GloVe (as semantic features), 32 boolean lexicon-based features from Recasens et al. (2013) and document vector representations (as contextual features) to distinguish between different uses of the same word. They find that when training a logistic regression classifier, the semantic features alone perform better than both the contextual and the combination of the two.

2.3 Classification Algorithms

Also performing bias classification at the sentence level, Vincze (2013) detects sentences containing weasel, hedge or peacock words from the WikiWeasel corpus with a precision of 0.74, recall of 0.69 and F_1 of 0.71, by using a dictionary lookup approach. Bellows (2018) reports an accuracy of

0.68 on a corpus of 2,143 biased sentences from news articles, vectorized using tf-idf and classified with a Mutlinomial Naive Bayes, and an accuracy of 0.77 for a CNN and 0.78 with a RNN. Finally, Hube and Fetahu (2018) achieve an F_1 measure of 0.70 using Random Forest on 686 manually annotated sentences from Conservapedia.

3 Dataset Description

We propose a procedure to semi-automatically derive a labeled corpus of biased sentences from a Wikipedia dump in any language, which, for this paper, we applied to the April 2019 dumps[2] for Bulgarian, French and English.

3.1 Tagset Curation

First, we manually compile a list of NPOV-related tags for each of the target languages using the names of relevant Wikipedia maintenance templates[3] ({{POV}}, {{NPOV}}, {{neutral point of view}}, {{peacock}}, etc.).

Most tags, however, vary in spelling, not only based on the context (e.g., inline or at the beginning of an article), but also because of the open and collaborative nature of Wikipedia. Table 1 shows the sixteen most frequent "weasel" tag variations, only five of which (in bold) are documented on Wikipedia. While the official tag is the most frequently used, the unofficial variations account for almost 35% of the most frequent ways to tag a page containing weasel words.

While it may be effortless for human editors to interpret the meaning of these variations, it is not trivial to automatically identify all NPOV-related ones. Simply extracting all the tags starting with the official form of "weasel" yields unrelated tags such as "weasel, back-striped" (an animal) or "weasel, ben" (a punk singer). For that reason, we automatically compiled exhaustive tag frequency lists in each language, and then manually selected the relevant variations of each.

3.2 Revision Extraction

We look for occurrences of the selected tags across all revisions of each page, going forward from the oldest one. When a biased revision is found, we follow its evolution until the POV tag disappears, at which point we assume the problematic content has been either rewritten or edited out. Next, we extract the tag together with the pair of adjacent revisions, where the older one is tagged as biased and the newer one is not. We opted for this diachronic retrieval method, rather than relying on the repertoire of articles in Wikipedia's "NPOV dispute" section (Herzig et al., 2011; Recasens et al., 2013) since the latter only features currently tagged articles, while our method digs NPOV violations from revision histories.

Tag	Count	Ratio
weasel	201,092	0.5748
weasel-inline	89,352	0.2554
weasel words	21,755	0.0622
weasel word	16,991	0.0486
weasel section	3,954	0.0113
weasel-section	3,743	0.0107
weasel inline	2,631	0.0075
weaselinline	2,213	0.0063
weasel-words	2,176	0.0062
weasel-word	2,102	0.0060
weaselword	1,967	0.0056
weasel-name	956	0.0027
weaselwords	503	0.0014
weasel_section	225	0.0007
weasel_words	124	0.0004
weasel_word	80	0.0002

Table 1: "Weasel" tag variation in English

3.3 Processing and Filtering

Each of these revision pairs undergoes a cleaning process using regular expressions to strip as much of the Wikipedia markup, links, and page references as possible, while preserving visible text and essential punctuation. At this point, we proceed to tokenize the text and split it into sentences using the rule-based tokenizer and sentencizer methods of spaCy (Honnibal and Montani, 2017), whose 2.1.3 version supports 51 languages. Finally, we replace all numbers with a special token (numtkn), strip all remaining punctuation, and convert everything to lowercase.

Our algorithm also extracts revision pairs where the second member was the subject of a redirect or vandalism, which we filter out. We then compare the revisions to obtain the lists of deleted and inserted sentences for each pair. In about 20% of the cases, the difference consists in simply deleting the NPOV tag, which we believe is an artifact of editorial wars (Sumi et al., 2011; Yasseri et al.,

[2] https://dumps.wikimedia.org

[3] For English, see https://en.wikipedia.org/wiki/Category:Neutrality_templates

2012), given the contentiousness of most NPOV-flagged topics. Another 20% of the revision differences we set aside are punctuation or case-related.

We further clear the dataset from outliers (mostly acts of vandalism) by removing those with more than 400 edited sentences. Finally, we exclude revision pairs with minor differences (character-based Levenshtein distance of 1), which are spelling corrections rather than bias resolution. Table 2 gives the number of initial, final and excluded revisions per language.

Revision pairs	BG	FR	EN
initial number	1,021	46,331	197,953
tag removal	–257	–10,255	–61,397
punct./case	–194	–5,967	–44,345
redir./vandalism	–56	–1,524	–17,154
deletions only	–33	–2,740	–11,331
insertions only	–28	–2,819	–2,938
spelling	–3	–136	–400
outliers	–2	–153	–609
Total pairs	448	22,737	59,779

Table 2: Number of revision pairs per language

To build the final corpora, we take all removed and added sentences (under 300 tokens) from the pre-filtered revisions for the positive and negative classes respectively. We balance the dataset by using unchanged sentences (also treated as negatives), as shown in Table 3.

Sentences	BG	FR	EN
Removed	4,756	105,939	800,191
Added	3,288	72,183	494,993
Unchanged	1,468	33,756	305,198
Total	9,512	211,878	1,600,382

Table 3: Number of sentences per language

4 Dataset Evaluation

Once we have collected the tagged/untagged revision pairs for each language (as per §3.2), we evaluate their potential for automatic bias detection. Our intuition is that the sentences that were removed together with the NPOV tag in the same edit likely contain some form of bias. Insertions, on the other hand, come with little guarantee of neutrality, so we focus on the removed sentences.

4.1 Protocol

For each language, we distribute the tagged/untagged revision pairs into four bins, based on the number of sentences that were removed in the edit (bin 1: 1 or 2 sentences removed, bin 2: 3–6, bin 3: 7–15, bin 4: 16 or more; these values were determined empirically to yield balanced bins in terms of revision pairs). Each annotator labeled 296 randomly picked sentences for a given language, distributed equally across the four bins. 72 of these sentences (24%) were shared by all annotators working on the same language, while the remaining 224 were labeled by a single annotator (cf. Table 4), thus allowing us to annotate more sentences while maintaining enough overlap to measure inter-annotator agreement (IAA). The Bulgarian sample was annotated by two native speakers, English by three with near-native proficiency, and French by two natives.

Lang	All	Ann1	Ann2	Ann3	Total
BG	72	224	224	—	520
FR	72	224	224	—	520
EN	72	224	224	224	744

Table 4: Number of sentences per annotator

The annotators were given identical instructions. For each sentence in their sample, they had to say whether it violated any of the NPOV principles stated in §1. The annotators were always presented with the full revision pair, so they had access to the context.

4.2 Dataset Evaluation Results

Since we had three annotators for English, we used Fleiss' κ to measure IAA. Tables 5 and 6 give the rate of positive annotations and IAA per language and per bin. On average, across all languages and bins, the annotators found 48% of positives in their samples, with an overall IAA of 0.41. Leaving out BG bin 4 (the only one with a negative κ), we get an average positive rate of 47% (std = 0.08) and an average κ value of 0.46 (std = 0.14). Our IAA coefficients are consistent with Vincze (2013), who had 200 English Wikipedia articles annotated by two linguists for weasel, peacock and hedge words, with IAA rates of 0.48, 0.45 and 0.46, respectively, and higher than the 0.35 reported by Hube and Fetahu (2018), who crowdsourced the annotation of sentences from Conservapedia into biased and unbiased. Identifying such phenomena

is thus not trivial but reasonable agreement can be expected.

Bin	BG	EN	FR	avg	std
1	0.34	0.51	0.47	0.44	0.07
2	0.64	0.45	0.45	0.52	0.09
3	0.63	0.45	0.38	0.48	0.11
4	0.63	0.52	0.34	0.50	0.12
avg	0.56	0.48	0.41	0.48	0.06
std	0.13	0.03	0.05	0.03	0.10

Table 5: Positives in annotations

Bin	BG	EN	FR	avg	std
1	0.32	0.55	0.67	0.51	0.15
2	0.22	0.58	0.44	0.41	0.15
3	0.32	0.31	0.61	0.41	0.14
4	−0.23	0.39	0.68	0.28	0.38
avg	0.16	0.46	0.60	0.41	0.18
std	0.23	0.11	0.10	0.08	0.21

Table 6: Inter-annotator agreement (Fleiss' κ)

About half of the annotated sentences turn out to be neutral. Below, we discuss the sources of the noise we have observed in our dataset (including the added sentences).

4.3 Sources of Noise

We identified two types of noise: pipeline-related and human-related. Pipeline-related noise is either noise introduced at the pre-processing phase (e.g., due to inconsistent sentence segmentation) or noise that remains despite our filtering and cleaning efforts (e.g., NPOV-unrelated edits longer than one character, differences resulting from the introduction of an infobox, differences consisting in changing the spelling of numbers).

Human editor-related noise comes from the data itself and stems from the behaviour of Wikipedia's editors. It includes edits which introduce bias (often intentionally, as in (1) below), vandalism, corrections of factual mistakes unrelated to bias, replacing bias with another bias (cf. (2)), and collateral edits, i.e., neutral sentences neighbouring biased ones indirectly targeted by a large-scope edit (cf. (3)). Below are some examples.

(1) a. *(before)* cardinal health inc is a holding company
b. *(after)* cardinal health is a healthcare company **dedicated to making healthcare safer and more productive**

(2) a. *(before)* its support is low only in the cholla province which has for nearly numtkn years supported kim dae jung a well known **leftist** politician born in that province who also served as president of south korea numtkn numtkn
b. *(after)* its support is low only in the jeolla province which has for nearly numtkn years supported kim dae jung a well known **progressive** politician born in that province who also served as president of south korea numtkn numtkn

(3) a. *(before)* from the numtkn th century confucianism was losing its influence on vietnamese society monetary economy began to develop but unfortunately in negative ways
b. *(after)* from the numtkn th century confucianism was losing its influence on vietnamese society and a monetary economy began to develop

5 Expressions of Bias

The manual annotation also highlighted the variety of bias expression. Previously, Recasens et al. (2013) had identified two major classes: epistemological and framing bias (subjective intensifiers and one-sided terms), where they considered the first one to group more implicit expressions such as factive and assertive verbs, entailment and hedges. Based on their work and Wikipedia's Manual of Style, we present biased examples from our corpus[4] and discuss them in terms of the overt/covert nature of the biased statement, its length (one or more words), and its level of ambiguity.

Subjective intensifiers are mostly expressed through single-word verbal and nominal modifiers (adverbs and adjectives) as in (4) and (5), but may also take the form of superlatives or quantifiers. They explicitly undermine tone neutrality by introducing overstatements and exaggerations (6).

(4) a. *(before)* some prominent liberals including scott reid were **strongly** critical of volpe s response
b. *(after)* some prominent liberals including scott reid **criticized** volpe s response

(5) *(before)* he is **truly** one of the greatest americans

(6) a. *(before)* this is an **absurd** statement because the cavalry of any age is designed first and foremost to run over the enemy and separate them as to make

[4]Examples are taken from the English evaluation subsets, where sentences are in lowercase, stripped of punctuation and numbers are replaced by `numtkn`.

them far more vulnerable to being overwhelmed and overrun

b. *(after)* this is **wrong** because the cavalry of any age is designed first and foremost to run over the enemy and separate them as to make them far more vulnerable to being overwhelmed and overrun

Clichés and jargon tend to be non-ambiguous but introduce low-frequency words in the corpus, as a result of being discouraged by Wikipedia.

(7) *(before)* x force was **concocted** by illustrator rob liefeld who started **penciling** the new mutants comic book in numtkn

Describing or analyzing rather than reporting events is a form of partiality harder to model, as it may not necessarily contain explicitly proscribed vocabulary.

(8) *(before)* he was a former club rugby and an opening batsman in club cricket but did not have the ability to make it all the way to the top level these two sports have become his particular area of expertise however he is very knowledgable on all sports that are played

(9) *(before)* however the most important consequence of the battle was that president lincoln was able to sieze upon the victory claim it as a strategic victory for the north and release his emancipation proclamation

Active voice may be used in cases like (10) to stress the agency of a participant in a situation, alongside a positively loaded support verb.

(10) a. *(before)* the united states department of justice indicted the company but **amway secured an acquittal**

b. *(after)* the united states department of justice indicted the company but **amway were acquitted**

To state a fact as an opinion is to use a weasel word to undermine the fact (11) or hide its source. While previous research shows the success of word-lists in detecting this particular type of bias (Recasens et al., 2013; Ganter and Strube, 2009), Vincze (2013) warns against the ambiguity of many of them. For example, *most* can be a weasel word (*Most agree that...*), a hedge (*most of his time*), a peacock (*the most touristic beach*) or neutral (*He did the most he could.*)

(11) a. *(before)* in the first invasion operation litani in numtkn the israeli military and south lebanon army sla occupied a narrow strip of land **ostensibly** as a security zone

b. *(after)* in the first operation litani in numtkn the israel defense forces and south lebanon army occupied a narrow strip of land **described** as the security zone

To state an opinion as a fact may be done with the use of an adverb (12) or an omission (13).

(12) a. *(before)* **in fact** the need for fast and secure fund transfers is growing and in the next year instant payments will quickly become the new normal for electronic fund transfers

b. *(after)* it is predicted that in the next year instant payments will become the standard for electronic fund transfers

(13) a. *(before)* in numtkn the journal won the praise of fascist leaders

b. *(after)* **there are some authors who retain** that the journal won the praise of fascist leaders

Intentional vagueness or the omission of factual information (14), is arguably the hardest type of bias expression to detect not only for machines, which are expected to recognize the lack of data as an informative feature, but also for humans, since filling factual gaps requires a fair amount of domain-specific knowledge.

(14) a. *(before)* as of numtkn it is the ethnic minority party in romania with representation in the romanian parliament

b. *(after)* as of numtkn it is the ethnic minority party in romania with representation in the romanian parliament **and is part of the governing coalition along with the justice and truth alliance and the conservatives**

6 Classification Experiments

The goal of the experiments is to assess the usefulness of the dataset in a sentence classification task. Our hypothesis is that having similar examples in both the biased and non-biased classes would help to single out discriminative words targeted by the NPOV-related edits.

Each dataset was split into a training set (80%), a development set (10%) on which we tuned the parameters, and a test set (10%) on which we ran a single evaluation with the best parameters.

6.1 Embeddings

We used fastText's classification function (Joulin et al., 2017), which implements a multinomial logistic regression algorithm on top of pretrained word embeddings. It uses word and character level embeddings to predict the class value of an instance. The parameter optimization was done by altering values for epoch (5, 10, 25), learning rate (0.1, 0.01, 0.05), word n-grams (1 to 5), minimum count (1–5), embedding dimensions (100, 300), loss function (softmax, ns, hs), minimum character level n-gram size (2, 3), using pretrained vectors or not, and learning rate update rate (50, 100).

When applying fastText's pretrained vectors,[5] we obtained comparable results for English and French without any significant gain, and with lower performance on Bulgarian. Thus, the final model chosen for its overall best performance across all three languages was trained without the use of an additional language model. The best performing values were then tried out on the test set.

6.2 Bag-of-Words Vectorization

We also experimented with classic bag-of-word vectorization with the stochastic gradient descent (SGD) (LeCun et al., 1998) and logistic regression (Hosmer and Lemeshow, 2000) algorithms. Each algorithm was run with the same settings on all three datasets to get the best average overall performances for precision, recall and F_1 measure. Parameter optimization was done using a grid search. Stop word lists were used for each language, which is the only language-specific aspect of the experiment.

The optimization for SGD ran 72 permutations with the following parameters:

- Bag-of-word n-gram size: unigrams only, unigrams and bigrams, unigrams to trigrams.
- Bag-of-word size: 100, 150, 300, 500, 1,000 and 3,000.
- Use idf reweighting or not.
- α value: 0.01, 0.001.

All the other parameters were set to their default values.[6] For logistic regression, 504 permutations were tested using the following settings:

- Same BOW n-gram size and BOW size and value type as SGD.
- C: 1.0e–3, 1.0e–2, 1.0e–1, 1.0e0, 1.0e+1, 1.0e+2 and 1.0e+3.
- Solver: sag, saga.

Using the training and development sets to run the grid search optimization on all three languages, the average F_1 measure was used to see which parameter values offered the best average performance across the board. The selected values were then used to run the same algorithm once on each language's training and test sets.

7 Results and Discussion

Table 7 shows the results for the experiments detailed in §6 for the SGD, fastText and logistic regression (LR) algorithms. For each performance measure, dataset section, algorithm and language, we provide results with respect to the *biased* class. The highest performance obtained on the test dataset of each language is in bold.

For the LR algorithm, the best performances were obtained using a C value of 0.001 with the saga solver using a unigram model of 100 features without inverse document frequency (idf) reweighting. The best parameters for the SGD used a model of unigrams to trigrams, with an α of 0.001 and idf reweighting. For fastText, the best performing parameter set used the default values[7] and a minimum of 5 occurrences per token.

Overall, the similar results between the development and test sets for each algorithm confirm that they did not overfit. Furthermore, all three measures have relatively low variance across languages, except for recall with SGD, which is considerably lower for Bulgarian (also impacting F_1) than for the other two languages.

We observe that FastText's vectorization and classification methods deliver higher precision upon larger datasets, but SGD and LR assure a much higher recall regardless of the number of examples.

While relatively better, the SGD performance level on the test set leaves room for improvement. This is likely due to the noise level in the sentences labeled as biased, which count many non-biased examples (see §4.2). The results are equally likely affected by the lexical and contextual ambiguity of the biased expressions, as discussed in §5. However, we do observe comparable best performance

[5] Available for 157 languages, pretrained on Common Crawl and Wikipedia (Grave et al., 2018) https://fasttext.cc/docs/en/crawl-vectors.html

[6] Version 0.21.2 of the sklearn toolkit.

[7] For version 0.8.3 of https://github.com/facebookresearch/fastText

Measure	Lang.	Dev-SGD	Test-SGD	Dev-fastText	Test-fastText	Dev-LR	Test-LR
Precision	BG	0.5387	**0.5886**	0.5324	0.5330	0.5182	0.5032
	FR	0.5059	0.5087	0.5533	**0.5520**	0.5151	0.5161
	EN	0.5112	0.5083	0.5656	**0.5634**	0.5230	0.5224
Recall	BG	0.4318	0.5049	0.4752	0.4937	0.6219	**0.6303**
	FR	0.8877	**0.8363**	0.5724	0.5721	0.6751	0.6739
	EN	0.8357	**0.8277**	0.5686	0.5718	0.5344	0.5354
F_1	BG	0.4794	0.5435	0.5022	0.5126	0.5653	**0.5596**
	FR	0.6444	**0.6146**	0.5627	0.5619	0.5844	0.5845
	EN	0.6334	**0.6291**	0.5671	0.5676	0.5286	0.5288

Table 7: Results for each language, dataset and classification method for the biased class

across corpora of varying size and languages from different families.

On the test set, our best overall average F_1 measure ranged between 0.56 and 0.62. This is lower than Vincze (2013)'s 0.71 or Hube and Fetahu (2018)'s 0.70, but our approach uses a large corpus, automatically derived from Wikipedia in any language with minimal language-specific input, applied to sentence-level bias detection, while Vincze (2013) used a monolingual, dictionary-based approach, and Hube and Fetahu (2018) relied on language-specific resources to extract multiple lexical and grammatical features. Our results set the baseline for sentence-level bias detection across the three languages of this corpus. Higher performance for a specific language may be achieved by a reconfiguration of the parameters or by the introduction of additional features.

8 Conclusion and Future Work

We presented a semi-automatic method to extract biased sentences from Wikipedia in Bulgarian, French and English. As this method does not rely on language-specific features, apart from the NPOV tag list and a stop word list, it can be easily applied to Wikipedia archives in other languages. It relies on the tags added by human editors in the articles that they considered biased. We retrieve the last tagged revision and the untagged revision following it and regard them respectively as biased and unbiased. By comparing the revisions, we get the lists of removed and added sentences.

We manually annotated 1,784 of the removed sentences, for all three languages combined, and found that only about half of them were actually biased. An average Fleiss' κ of 0.41 (0.46 if ignoring an outlier), consistent with the literature, indicates that the task is not trivial even for humans.

Using our corpora, we tested three classification algorithms: bag-of-word vectorization with SGD, fastText, and logistic regression.

In future work, we would like to improve the quality of the dataset by addressing issues uncovered during the human evaluation, such as incoherent sentence segmentation, enumerations, minor edits and remaining noise. Another conceivable optimization is to segment the dataset into two or more subsets according to the main forms of bias expression (e.g., explicit vs implicit). It would allow to explore and evaluate different forms of bias separately, which in turn might motivate differential classification techniques. Finally, populating the negative examples class with sentences from Wikipedia's Featured Articles (in line with Bhosale et al. 2013) might help reduce class ambiguity by reinforcing the contrast between neutral encyclopedic tone and expressions of bias.

Acknowledgments

This work has been supported by the Ministère de l'Économie et de l'Innovation du Québec (MEI). We would like to thank the annotators for their help with the quality evaluation process and the anonymous reviewers for their insightful comments.

References

Khalid Al Khatib, Hinrich Schütze, and Cathleen Kantner. 2012. Automatic detection of point of view differences in Wikipedia. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 33–50.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA), Valletta, Malta.

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM* 61(6):54–61.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *EMNLP-2018* .

Martha Bellows. 2018. *Exploration of Classifying Sentence Bias in News Articles with Machine Learning Models*. Ph.D. thesis, University of Rhode Island.

Shruti Bhosale, Heath Vinicombe, and Raymond Mooney. 2013. Detecting promotional content in wikipedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pages 1851–1857.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *J. Commun.* 57(1):163–173.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task*. Association for Computational Linguistics, pages 1–12.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. pages 173–176.

Njagi Dennis Gitari, Zuping Zhang, Hanyurwimfura Damien, and Jun Long. 2015. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering* 10(4):215–230.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Livnat Herzig, Alex Nunes, and Batia Snir. 2011. An annotation scheme for automated bias detection in wikipedia. In *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, Stroudsburg, PA, USA, LAW V '11, pages 47–55.

Nicholas P Hirning, Andy Chen, and Shreya Shankar. 2017. Detecting and identifying bias-heavy sentences in news articles. Technical report, Stanford University.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear* .

David W. Hosmer and Stanley Lemeshow. 2000. *Applied logistic regression*. John Wiley and Sons.

Christoph Hube and Besnik Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, pages 1779–1786.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1113–1122.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 427–431.

Sicong Kuang and Brian D Davison. 2016. Semantic and context-aware linguistic model for bias detection. In *Proc. of the Natural Language Processing meets Journalism IJCAI-16 Workshop*. pages 57–62.

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. 1998. Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*. Springer-Verlag, London, UK, UK, pages 9–50.

Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. pages 1153–1161.

Gabriel Murray and Giuseppe Carenini. 2009. Detecting subjectivity in multiparty speech. In *Tenth Annual Conference of the International Speech Communication Association*.

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Umashanthi Pavalanathan, Xiaochuang Han, and Jacob Eisenstein. 2018. Mind your POV: Convergence of articles and editors towards wikipedia's neutrality norm. *Proc. ACM Hum. -Comput. Interact.* 2(CSCW):137:1–137:23.

Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On measuring bias in online information. *ACM SIGMOD Record* 46(4):16–21.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1650–1659.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*.

Lee D Ross, Teresa M Amabile, and Julia L Steinmetz. 1977. Social roles, social control, and biases in social-perception processes. *J. Pers. Soc. Psychol.* 35(7):485.

Nicole Shadowen. 2019. Ethics and bias in machine learning: A technical study of what makes us "good". In Newton Lee, editor, *The Transhumanism Handbook*, Springer International Publishing, Cham, pages 247–261.

R Sumi, T Yasseri, A Rung, A Kornai, and J Kertesz. 2011. Edit wars in wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. ieeexplore.ieee.org, pages 724–727.

A Tversky and D Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Veronika Vincze. 2013. Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. pages 383–391.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International conference on intelligent text processing and computational linguistics*. Springer, pages 486–497.

Theresa Wilson and Stephan Raaijmakers. 2008. Comparing word, character, and phoneme n-grams for subjective utterance recognition. In *Ninth Annual Conference of the International Speech Communication Association*.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, Stroudsburg, PA, USA, CSLDAMT '10, pages 152–158.

Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. 2012. Dynamics of conflicts in wikipedia. *PLoS One* 7(6):e38869.