Mitigation of Unintended Biases against Non-Native English Texts in Sentiment Analysis

Alina Zhiltsova[1], Simon Caton[2], and Catherine Mulwa[1]

[1] School of Computing, National College of Ireland, Ireland
x18121900@student.ncirl.ie, alinazhiltsova@gmail.com,
catherine.mulwa@ncirl.ie
[2] School of Computer Science and Informatics, University College Dublin, Ireland
simon.caton@ucd.ie

Abstract. Currently the demand for text analytics grows due to the fact that textual data is being created online in large amounts. A number of tools are available for various tasks related to natural language processing such as sentiment analysis and text pre-processing. The majority of these tools are trained using unrepresentative data, which may lead to unintended biases altering the results. Previous research indicates that sentiment analysis tools show gender and race biases, and word embeddings discriminate against women. This research investigates previously undefined non-native speaker bias in sentiment analysis, i.e. unintended discrimination against English texts written by non-native speakers of English. Non-native speakers of English tend to use cognates, English words that have origin in the speaker's language. To measure the non-native speaker bias in 4 lexicon-based sentiment analysis systems, a new Cognate Equity Evaluation Corpus was created, based on previous work in the literature for measuring racial and gender biases. The tools gave significantly different scores to English texts with features of non-native speakers. The bias discovered in lexicon-based tools was mitigated by updating 4 lexicons for English cognates in 3 languages. This paper proposes a generalisable framework for measuring and mitigating non-native speaker bias.

Keywords: Fairness in Machine Learning, Natural Language Processing, Bias Mitigation, Non-Native Speaker Bias, Sentiment Analysis

1 Introduction

The number of domains depending on data analysis and machine learning has been growing with a previously unseen pace. As it happens, the fairness of machine learning models and their application is becoming increasingly more important. There is, however, no single established definition of unintended bias [5]. In this paper, bias is understood as a preference of some attributes over others. When this preference is not intentional, the fairness of the model needs to be questioned [1]. Unfairness in machine learning can lead to consequences of various scales including not hiring a person because of their gender that was disfavoured during an automated CV screening, or not allowing a prisoner parole because of their skin colour rather than their criminal record and/or recent reform (ibid.).

Natural language processing (NLP) is an area of machine learning which includes various types of analysis of textual data, e.g. text classification, sentiment analysis, text generation and so on. Just like in machine learning in general, unintended biases have been noticed in NLP - e.g. word embeddings which are used for various tasks ranging from Google news sorting to abusive content filtering have been found to be biased against women [3]. It has also been found that sentiment analysis systems which are used both in academia and industry can discriminate against races and genders [14]. Others have also observed that gender can significantly affect results [7, 8, 15]. Yet, there are few approaches for the general study, identification or mitigation of these effects.

It has been shown that non-native speakers of English tend to use cognates, i.e. English words that look similar to the words of their native language in their English texts more often (e.g. French speakers are more likely to use 'fatigue' (French/Latin origin, cognate of French 'fatigué'), and German speakers are more likely to use 'weary' (Proto-Germanic origin, cognate of Frisian 'wuurig') in the same context because of the etymology of those words [16]). Since the majority of sentiment analysis systems are trained on native speaker data, we suggest that sentiment analysis systems can be biased against non-native speaker texts leading to discrimination against non-native speakers of English. This paper provides a methodology for measurement and mitigation of this bias.

To address the issue of unintended biases the following research question was specified and tackled:
'To what extent can the expansion of the Equity Evaluation Corpus (EEC) with cognates enable measuring and, where necessary, mitigating bias in lexicon-based sentiment analysis systems against English texts written by non-native speakers (speakers of French, Italian and Spanish)?'

To address the research question several sentiment analysis (VADER, SentimentR, TextBlob, Afinn) systems were measured and debiased. The specified research question aims at solving the issues of unintended biases in natural language processing, and bringing value to NLP researchers, linguists, and data analysts who deal with textual data. The results of this research will help practitioners reduce a method bias in the application of NLP methods. This will especially be the case in the use of online data, for example from Social Media, online reviews, blogs etc. The sentiment analysis systems specified were chosen based on their popularity and functionality. The languages are chosen based on comparatively large number of English cognates they have.

The rest of the paper is structured as follows: section 2 presents relevant related work and its influence on this work; section 3 describes the methodology applied; section 4 summarises the implementation, evaluation and results of our study; finally, section 5 concludes the paper.

2 Related Work

Fairness in machine learning is a growing concern both in industry and academia. The scope of the reviewed literature is from 1996 to 2019.

2.1 Fairness and Unintended Bias in Machine Learning

As the reliance on machine learning for decision making grows in various spheres of business and life, the fairness of those decisions becomes increasingly important. To the best of the authors' knowledge there is no standardised definition of the concept of fair machine learning, and depending on the domain algorithmic fairness can be understood in a number of ways. It is suggested, that machine learning algorithms or models have to be biased towards some particular decision, however, when this bias is unintended, a model or algorithm can be considered unfair [1].

The source of bias can be different on a case by case basis, but generally it is observed that bias comes from the training data [5] - in one example of image recognition, people with darker skin were proved to be more difficult to identify than people with fairer skin, because the training dataset was imbalanced and contained significantly more images of white people [4]. Dataset imbalance is also observed in NLP, in toxicity (i.e. toxic language) and abuse detection in particular [15]. Unintended bias can be introduced to the data by personal biases during manual data labelling [8]. Another source of bias can sometimes be model architecture [15] - the models that see each word as a separate feature can show higher rate of false positive scores in abusive content classification.

Generally, there are 2 types of fairness that provide guidelines for determining a case-specific approach - *individual* fairness and *counterfactual* one. Individual fairness comes from the classification tasks - it is suggested that two similar observations or data points should be classified into the same category. Originally it was achieved using a distance metric that would calculate how similar the observations are [9]. Counterfactual fairness depends on the concept of adding counterfactuals for each level of data such as sex or skin colour [13].

In this paper, individual fairness is assumed to be the guiding type of fairness - i.e. it is suggested that similar sentences, one written by native speakers and another written by non-native speakers should get similar results from NLP systems. However, unlike in [8], in this case it is not connected with classification, and applies to a sentiment analysis score instead. The following section offers more details on state-of-the-art bias measurement and mitigation in textual data.

2.2 Biases in Natural Language Processing

Existing Approaches to Measuring Biases in NLP:

When it comes to NLP, word replacement as a method to determine if bias is prevalent. To measure the bias against identity terms such as ‘muslim’, ‘gay’, and others in toxicity classification, [8] have created a synthetic testing dataset which contained more than 70000 sentences, half of which were toxic and half of which were not. The dataset was based on several manually created templates such as 'IDENTITY people are just like everyone else'. Using replacement, identity terms combined with the templates produced the final dataset used in the publication.

A similar approach was used for measuring bias against age in sentiment analysis. Sentences containing explicit ('old', 'young') or implicit (words that are likely to be used by old or young people) age encodings were used to test 15 sentiment analysis tools. The sentiment was measured for sentences with the term originally contained in the sentence, and then with the opposite one [7].

[14] propose to use Equity Evaluation Corpus[3] the creation of which is described in detail in that paper. The authors created a corpus of sentences using templates they suggested for various emotions and for the neutral sentences. The emotions included anger, fear, joy and sadness, based on the Roget's Thesaurus. The sentence templates used in this paper were based on these same four emotions.

The three examples described above are very useful for our approach, as they focus on the influence of a particular word like an identity term or an age related adjective on the outcome of the sentiment analysis results. The focus of this paper is also word-based - it relies on synsets, or pairs of words, that include a cognate and its non-cognate synonym.

Mitigation of Unintended Biases in Natural Language Processing:

It is found, that generally mitigation of biases in sentiment analysis and related tasks can be performed either through changing an algorithm (in-processing) or through changing the training data (pre-processing).

Changing the training data was observed in several papers related to sentiment analysis. To mitigate the bias related to age, 2 custom sentiment analysis tools based on the bag-of-words text classification were built with a logistic regression. Training data used for *Sentiment140*[4] was modified to suit the needs of the research - i.e. measure bias against age originating from the training data. Training data for the dataset was filtered to identify records containing words related to age like 'young', 'old' and some others resulting in about 14 000 tweets. As a result it was confirmed that the age bias in sentiment analysis at least partially derives from the labels in the training data, and as long as the examples including words related to age are excluded, bias can be significantly mitigated [7].

Similarly to the previous example, mitigation of biases can also be done through an unsupervised balancing of the training dataset [8].

It is also important to point out that in some cases mitigation of biases is not attempted due to the complexity of the task. [14] measured the bias against race and gender for 200 sentiment analysis tools, however they concluded that removing the bias is better left for the future work.

[3] https://saifmohammad.com/WebPages/Biases-SA.html, source of the data, accessed on the 17th of July, 2019.

[4] http://www.sentiment140.com/, accessed on the 22nd of July, 2019.

2.3 Existing Problems in NLP Connected to Non-Standard English

Since the main focus of this paper is on English texts produced by non-native speakers, it is important to understand how different this type of English texts can be and how strongly it influences the outcomes of various NLP tasks.

The term *non-standard English* can be applied to all the non-standard dialects and variants of English spoken by non-native speakers, and also to the non-standard Englishes spoken by native speakers such as Cockney accent or Scottish accent. Each non-native speaker might use a slightly different version of English depending on their native language and level of proficiency in English [17].

One of the features of non-native Englishes includes using cognates - English words that have a similarly looking or sounding word in the speaker's native language. It was observed that even speakers with high level of English proficiency tend to prefer cognates over other words if a cognate exists. The presence of a cognate word in a piece of text can be enough to identify a speaker's native language. Examples of such words include a pair 'weariness' (German origin) - 'fatigue' (French origin) - a German speaker is more likely to use the first word, whereas a French speaker is likely to choose the second one because of the etymology of the words ([16]). This observation is the basis of the project's hypothesis - since speakers of different languages use different cognates in the same situations, their texts might get different sentiment analysis results where no difference was intended.

Non-standard English includes not only different vocabulary choices, but also different pronunciation. Various accents of English introduce a great variety of sounds for the same letter combinations. This can lead to issues in automated speech recognition (ASR) [17], and to transfer of pronunciation into spelling [2].

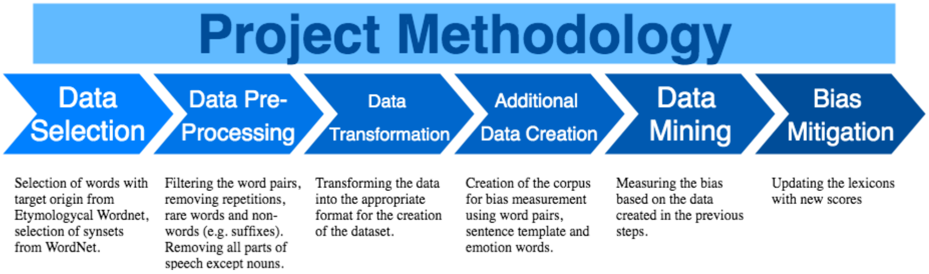
Consideration of these and other specifics of non-standard English is essential for NLP tasks, due to the fact that English is the primary communication language in many online communities and geographical locations regardless of the native language of the speakers.

3 Methods

Both components of the adopted scientific methodology - measurement and mitigation of biases - are implemented following a modified KDD approach [10]. Figure 1 presents the 6 phases which were followed during the implementation. Unlike the traditional KDD approach, it included the phase of additional data creation i.e. a dataset for measuring the bias that was compiled using the available word pairs and sentence patterns. Another addition is the last phase of bias mitigation that is normally not a part of a KDD method.

3.1 Data

For the implementation of this work, 5 data sources were used, 4 of which are open source, and one created by the authors. To process these data, 4 sentiment analysis lexicons (VADER, SentimentR, TextBlob, Afinn) were used.



Selection of words with target origin from Etymologycal Wordnet, selection of synsets from WordNet.

Filtering the word pairs, removing repetitions, rare words and non-words (e.g. suffixes). Removing all parts of speech except nouns.

Transforming the data into the appropriate format for the creation of the dataset.

Creation of the corpus for bias measurement using word pairs, sentence template and emotion words.

Measuring the bias based on the data created in the previous steps.

Updating the lexicons with new scores

Fig. 1. Sentiment Analysis Bias Measurement and Mitigation Methodology

Etymological WordNet: Etymological WordNet [6] is a collection of etymological origins of more than 500 000 words and more than 6 million links. The information for the dataset was collected from Wiktionary in 2013 and made available in 2014. The choice was motivated by related work presented in section 2 [16]. In total the dataset contains 3 columns of 6 000 000 rows.

WordNet: WordNet [11] (also available as a package in Python) is similar to a thesaurus - a collection of synonyms, antonyms, and other word groups. Its functionality is used to collect synsets - sets of synonyms - for English words that have a specific origin.

Reddit L2 Dataset: Reddit L2 Dataset [16] is a collection of Reddit posts in English written by French speakers, and was used to find French cognates that are in use. Nearly 90 000 posts were scanned for the cognates. Cognate lists of other languages did not require this step due to low numbers that could bias the results of statistical testing.

Equity Evaluation Corpus: Equity Evaluation Corpus is a dataset created by [14] to measure the bias against gender and race in sentiment analysis. The corpus has 8640 sentences created using 40 emotion words and 11 sentence patterns.

Creation of The Bias Measurement Dataset The Cognate Equity Evaluation Corpus (CEEC) is based on the Equity Evaluation Corpus (EEC) [14] created to measure gender and racial biases.

To provide a reliable corpus for non-native speaker bias measurement, a sentence template, EEC emotion words and cognate pair words were used.

The template used is ‘Talking about OBJECT made me feel EMOTION WORD.’. This template is based on the template from EEC ‘PERSON SUBJECT made me feel EMOTION WORD.’ There are several reasons behind not introducing more sentence templates. First is that sentence templates require sufficient time to be verified, as there are multiple methodologies behind creating them, and thus left for future work. Second is that the purpose of the corpus is to provide NLP researchers with an efficient way of detecting non-native speaker biases. Assigning sentiment takes different amount of time depending on the sentiment tool and the amount of text. Since the number of word pairs is much higher than in EEC, it was decided that one template with 4 emotions and 40 emotion words would be sufficient at this stage.

In every sentence the OBJECT was replaced with one of the words from each cognate pair. To get a cognate pair, all the English words with origin in the target language were selected from the Etymological Wordnet. Then each word was assigned a synset (synonym set) using the WordNet package in Python. Every word had a different number of synonyms, ordered by how often they are used. Only the most commonly used synset was kept for each word to scope the study and ensure balance in the dataset. Only noun pairs were kept for 2 reasons: 1) every part of speech would require a different template, so it was decided to keep only one due to the time limits; 2) nouns were the most represented part of speech. Different numbers of pairs for each origin language correspond to the nature of etymology of English words.

The emotion words for the template were retrieved from the EEC. In total 4 emotions were present in EEC - anger, fear, joy, sadness. Each emotion was represented by 10 words, e.g. ‘angry’, ‘frustrated’, ‘serious’, ‘enraged’, ‘excited’, ‘relieved’.

In total 3 versions of CEEC were created, as can be seen in Table 1.

Table 1. Cognate Equity Evaluation Corpus

Language	Cognate Pairs	Emotion Words	Total Number of Sentences
French	292	40	23360
Spanish	32	40	2560
Italian	45	40	3600

3.2 Feature Selection

The features required for this work are the “level” of the sentiment analysis and the text that will be used. When it comes to the level, sentence is chosen as a lot of textual data is being created online in microblog environments such as Twitter or Facebook, and the texts of this kind are usually short and do not always include more than 2-3 sentences. Another reason is the field standard - as was observed in the literature, most papers focus on sentences.

Sentiment analysis results may differ depending on capitalisation, punctuation marks, word choice, contrasting conjunctions (e.g. ‘but’), negators (e.g. ‘not’), degree modifiers (e.g. ‘very’, ‘slightly’, ‘a lot’) [12]. Since the focus of the paper is to identify and mitigate discrimination against non-native speaker texts that include cognates, we focus only on one feature: word choice, and examine its influence on the outcome (i.e. observed sentiment).

Languages of word origin chosen for this paper are French, Spanish and Italian. The choice is motivated by a high number of cognates in these languages due to etymological links between English and these languages.

3.3 Sentiment Analysis Systems Used

The work described in this paper relies on freely available sentiment analysis tools that can be seen in Table 2. To be chosen for the work, the tool had to satisfy the following requirements: it has to be completely free or have a free trial version; it has to be available for English language texts; it has to be available whether in R or Python, the languages commonly used by data scientists and NLP researchers; the sentiment value assigned to sentences has to be numeric and not categorical (e.g. 'good', 'bad', 'neutral'). While the selected tools are not the only ones that satisfy these requirements, they are all used in a variety of settings (e.g. teaching and research) and diverse enough to afford some degree of variability in the results.

Table 2. Sentiment Analysis Systems Used

System	Lexicon Size
VADER [12]	7517
Afinn[5]	3382
TextBlob[6]	2930
SentimentR[7]	11710

3.4 Measuring Bias in 4 Lexicon Based Sentiment Analysis Tools

To measure bias the CEEC dataset created for this work was used. Sentiment score was assigned to all the sentences using the 4 sentiment analysis tools - VADER, Afinn, and TextBlob in Python, and SentimentR in R. The sentiment scores from all tools were merged into one dataset, and the differences were calculated for a more detailed overview.

Assignment of the sentiment was performed using special commands available in the libraries. The standard settings were used for all the tools to achieve the score that is embedded in the lexicons of each tool.

4 Evaluation and Results

4.1 Evaluation and Results before Bias Mitigation

To evaluate the results, a Wilcoxon signed rank test was performed for each emotion. For each origin language and for each emotion the group of sentences containing the cognate words was compared to the group of sentences with no cognate word. Where the difference was significant at p-value less than 0.05, the presence of bias was confirmed. For each sentiment system the results were different. The scores were recorded in tables for each of the cognate origin language - French (Table 3), Spanish (Table 4), and Italian. Only the first two are presented as there was no significant bias observed against Italian cognates. SentimentR shows bias against the biggest number of French cognates - 133 out of 292. TextBlob shows bias against the least number of cognates - 19 for polarity and 20 for subjectivity. Essentially this means, that when a French person uses a cognate, e.g. ‘deficit’ which has a French etymology instead of ‘shortage’, the resulting sentence is likely to receive a more extreme sentiment score. It can be seen that the bias against cognates is present in sentences with various emotions in almost all sentiment tools. When it comes to Spanish cognates, SentimentR also shows bias against the biggest number of cognates - 6 out of 32. V=0 indicates that the scores for sentences with cognates and sentences without cognates were identical and as a consequence no bias was observed, as in the case of TextBlob. The scores were checked to ensure this. Measurement of bias against Italian cognates showed no significant bias due to the fact that none of the words in the pairs is present in any of the default lexicons.

Table 3. Results of Significance Testing for French Cognates

	Tool	Emotion	SignDiff	P_value
1	Vader	anger	yes	0.01541
2	Tblob_Pol	anger	yes	0.004203
3	Tblob_Subj	anger	no	0.2802
4	Afinn	anger	yes	0.01114
5	SentR	anger	yes	8.448e-06
6	Vader	fear	yes	0.009291
7	Tblob_Pol	fear	yes	0.0001365
8	Tblob_Subj	fear	no	0.7473
9	Afinn	fear	yes	0.01114
10	SentR	fear	yes	5.645e-06
11	Vader	joy	no	0.5923
12	Tblob_Pol	joy	yes	1.815e-05
13	Tblob_Subj	joy	yes	0.02074
14	Afinn	joy	yes	0.01114
15	SentR	joy	yes	0.03043
16	Vader	sadness	yes	0.01857
17	Tblob_Pol	sadness	yes	0.002438
18	Tblob_Subj	sadness	no	0.2796
19	Afinn	sadness	yes	0.01114
20	SentR	sadness	yes	8.419e-06

Table 4. Results of Significance Testing for Spanish Cognates

	Tool	Emotion	SignDiff	P_value
1	Vader	anger	no	0.06409
2	Tblob_Pol	anger	no	V=0
3	Tblob_Subj	anger	no	V=0
4	Afinn	anger	yes	0.001904
5	SentR	anger	yes	0.0245
6	Vader	fear	no	0.06451
7	Tblob_Pol	fear	no	V=0
8	Tblob_Subj	fear	no	V=0
9	Afinn	fear	yes	0.001904
10	SentR	fear	yes	0.0245
11	Vader	joy	no	0.06456
12	Tblob_Pol	joy	no	V=0
13	Tblob_Subj	joy	no	V=0
14	Afinn	joy	yes	0.001904
15	SentR	joy	yes	0.02469
16	Vader	sadness	no	0.06458
17	Tblob_Pol	sadness	no	V=0
18	Tblob_Subj	sadness	no	V=0
19	Afinn	sadness	yes	0.001904
20	SentR	sadness	yes	0.02452

Mitigating the Bias: To mitigate the bias against texts written by non-native speakers of English, the steps described below were taken. Due to comparatively smaller sample of sentences representing positive emotions (negative to positive sentences ratio is 3:1) only debiasing tools for negative emotions was added as objective of the research, to ensure fair and generalisable results.

First, the results of the bias measurement were observed for every sentiment system, and the word pairs with difference in scores were organised into lists. The lists were updated in the following way - cognates that had different score in the lexicon were changed to have the same score as the non-cognate word, and the cognates that were not in the lexicons were added with the score equal to the non-cognate word from the pair. This was done by assigning the scores to the words, and then adding the resulting cognate lists.

For Vader there is a special command that was used to update the lexicon by adding the words with their scores.

For Afinn no such command was found. To update the lexicon, the TXT file was found in the Python directory, and amended according to the bias. This file then can be used instead of the original lexicon, by either using a special set of commands, or by replacing the original file with the updated lexicon.

TextBlob debiasing is similar to Afinn in terms of manually updating the lexicon file in the Python directory. The file has to have the same name as the original lexicon. In the case of different lists for different languages it can create some inconvenience, however, renaming files is not time consuming and guarantees that the approach will work.

SentimentR allows adding any lexicon to the system. To debias the tool, the original lexicon was amended to reflect the correct scores for the cognates and saved as a new lexicon key. Unlike all the other tools, SentimentR allows to have various lexicons at the same time - i.e. lexicons debiased for French, and Italian cognates can exist as separate files simultaneously.

4.2 Evaluation and Results after Bias Mitigation

To determine the effectiveness of the debiasing process, the second round of significance testing was performed. The results show that after debiasing, the sentences with Spanish cognates and sentences without them receive identical scores in all 4 sentiment analysis systems for not only 3 negative emotions, but also the positive one (joy).

As for the French cognates, it was possible to debias all sentiment tools for all negative emotions - sadness, anger, and fear.

The test results show that SentimentR and TextBlob still show bias against cognates in sentences with positive emotion words related to joy. Debiasing tools for positive emotions is not in the objectives of this research, however, some insights given below were discovered.

After analysing the pairs with differences in scores it was concluded that in case of SentimentR the bias remained where synonyms consisted of 2 words, e.g. for French cognate ‘boutique’ the paired synonym is ‘dress shop’. SentimentR is structured in such a way that phrases like this get broken down and receive different scores. In the case of TextBlob, the reason behind the remaining bias is that the lexicon has different scores for various meanings of the affected words. To debias further, additional filtering and more data are necessary. It is also identified, that due to the fact that emotion words related to negative emotions such as anger, fear, and sadness are more extreme in scores, the bias for them was mitigated without additional filtering or data.

The approach implemented in this paper can be further generalised and used to debias sentiment analysis systems for cognates not only in English but other languages as well, as can be seen in Figure 2. It was successfully performed for 4 emotions in 4 tools for Spanish cognates, and for 3 emotions in 4 tools for French cognates.

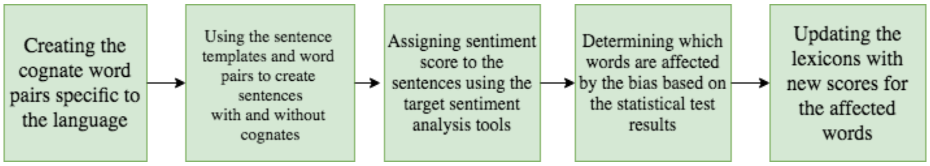

Creating the cognate word pairs specific to the language
Using the sentence templates and word pairs to create sentences with and without cognates
Assigning sentiment score to the sentences using the target sentiment analysis tools
Determining which words are affected by the bias based on the statistical test results
Updating the lexicons with new scores for the affected words

Fig. 2. Generalisable Bias Measurement and Mitigation Framework

5 Conclusion

In this paper, we have illustrated that texts with non-native speakers characteristics are likely to be handled quite differently by sentiment analysis tools. To show this, we expanded the Equity Evaluation Corpus (EEC) to enable the measurement of bias against English texts written by speakers of French, Italian, and Spanish in lexicon-based analysis systems (VADER, SentimentR, TextBlob, Afinn) for 4 emotions. We then demonstrated how to mitigate this bias against speakers of Spanish and French in 4 sentiment systems for 3 emotions. No bias was detected against speakers of Italian, so no mitigation was required.

This work can be extended in a number of ways. Increasing the number of languages observed and the number of sentence patterns used would permit a more expansive discussion of non-native speaker bias. It would be useful to also observe the bias against more emotions, especially positive ones. We have studied dictionary-based tools, as such corpus based merit study too, but would require extensions to be methodology developed here as it would involve building machine learning models to facilitate sentiment analysis of texts.

The results of this research contribute to fairness in machine learning literature, ensuring that non-native speakers of English face less discrimination in research designs. The findings of this work can enable data analysts and researchers improve the way textual data is analysed to avoid unintended bias against English texts produced by non-native speakers at a bigger scale.

References

1. Bantilan, N.: Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. Journal of Technology in Human Services 36(1), 15–30 (2018)
2. Blodgett, S.L., O'Connor, B.: Racial disparity in natural language processing: A case study of social media African-American English. arXiv preprint arXiv:1707.00061 (2017)

3. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 4349–4357. Curran Associates, Inc. (2016)
4. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency. pp. 77–91 (2018)
5. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 797–806. ACM (2017)
6. De Melo, G.: Etymological wordnet: Tracing the history of words. In: LREC. pp. 1148–1154. Citeseer (2014)
7. Díaz, M., Johnson, I., Lazar, A., Piper, A.M., Gergle, D.: Addressing age-related bias in sentiment analysis. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. p. 412. ACM (2018)
8. Dixon, L., Li, J., Sorensen, J., Thain, N., Vasserman, L.: Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 67–73. ACM (2018)
9. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226. ACM (2012)
10. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39(11), 27–34 (1996)
11. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, MA (1998)
12. Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media (2014)
13. Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: Advances in Neural Information Processing Systems. pp. 656–666 (2017)
14. Kiritchenko, S., Mohammad, S.: Examining gender and race bias in two hundred sentiment analysis systems pp. 43–53 (Jun 2018)
15. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2799–2804 (2018)
16. Rabinovich, E., Tsvetkov, Y., Wintner, S.: Native language cognate effects on second language lexical choice. Transactions of the Association for Computational Linguistics 6, 329–342 (2018)
17. Tatman, R.: Gender and dialect bias in Youtube's automatic captions. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. pp. 53–59 (2017)