

Bias-Aware Lexicon-Based Sentiment Analysis

Mohsin Iqbal
Information Technology
University of the Punjab,
Pakistan
mi308@itu.edu.pk

Asim Karim
Computer Science, SBASSE,
Lahore University of
Management Sciences,
Pakistan
akarim@lums.edu.pk

Faisal Kamiran
Information Technology
University of the Punjab,
Pakistan
faisal.kamiran@itu.edu.pk

ABSTRACT

Sentiment analysis of textual content is widely used for automatic summarization of opinions and sentiments expressed by people. With the growing popularity of social media and user-generated content, efficient and effective sentiment analysis is critical to businesses and governments. Lexicon-based methods provide efficiency through their manually developed affective word lists and valence values. However, the predictions of such methods can be biased towards positive or negative polarity thus distorting the analysis. In this paper, we propose Bias-Aware Thresholding (BAT), an approach that can be combined with any lexicon-based method to make it bias-aware. BAT is motivated from cost-sensitive learning where the prediction threshold is changed to reduce prediction error bias. We formally define bias in polarity predictions and present a measure for quantifying it. We evaluate BAT in combination with AFINN and SentiStrength – two popular lexicon-based methods – on seven real-world datasets. The results show that bias reduces smoothly with an increase in the absolute value of the threshold, and accuracy increases as well in most cases. We demonstrate that the threshold can be learned reliably from a very small number of labeled examples, and supervised classifiers learned on such small datasets produce poorer bias and accuracy performances.

Categories and Subject Descriptors

H.4 [Data Mining]: Text Mining

Keywords

Sentiment Analysis, Bias, Lexicon-Based Methods, Discrimination Aware Data Mining, Fairness in Learning

1. INTRODUCTION

Sentiment analysis is an important task in text mining that extracts and summarizes opinions and sentiments from

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'15 April 13-17, 2015, Salamanca, Spain.

Copyright 2015 ACM 978-1-4503-3196-8/15/04...\$15.00.

<http://dx.doi.org/10.1145/2695664.2695759>

unstructured textual content. It is used widely in practice to analyze reviews, blogs, and social media for opinions and sentiments expressed on products, services, events, organizations, and individuals. The most common form of sentiment analysis is polarity detection, i.e., determining whether the sentiment expressed in a document is positive or negative. This task has broad applications ranging from product/service assessment to political opinion analysis; its adopters are relying upon it for accurate actionable knowledge in their respective domains [12][17]. For instance, nowadays it is normal to look at reviews before buying a new product or watching a new movie. It is therefore desirable to have an automatic sentiment analysis method that processes the reviews and produces a polarity report.

Sentiment analysis and polarity detection methods can be supervised or unsupervised in nature. Supervised methods rely upon sufficient quantities of labeled data to learn a model like a naive Bayes classifier. Creating sufficient labeled data for the diverse applications and domains is generally practically infeasible. Unsupervised methods do not require labeled data and rely upon a predefined sentiment lexicon or affective word list to predict the polarity of documents. Thus, these methods have become very popular in recent years. In [6] eight state-of-the-art lexicon-based methods, e.g., SentiStrength [16], SentiWordNet [5], SASA [18], and SenticNet [3], are analyzed on real-world datasets, e.g., Web 2.0 labeled dataset [15], with strong performances.

Recently, it has also been shown that many lexicon-based sentiment analysis methods are biased towards positive polarity [6]. Figure 1 (taken from [6]) shows prediction polarity ratios of different methods on different datasets. The y-axis shows the difference between the percentage of positive and negative messages (as predicted by the methods and represented by different lines) and the x-axis identifies the datasets. The red dotted line represents the ground truth for the datasets. From Figure 1, it is clear that all methods produce biased results that deviate from the ground truth significantly (generally towards positive polarity). Such a systematic bias in prediction is highly undesirable for practical sentiment analysis as it can produce misleading decisions. Thus, despite the ease of use of lexicon-based methods they suffer from this major problem of biased predictions.

In this paper, we address the problem of bias in lexicon-based sentiment analysis methods by presenting BAT (Bias-Aware Thresholding). BAT can be combined with any lexicon-based method to make it bias-aware. Motivated from cost-sensitive learning [4], BAT involves changing the prediction threshold in a manner that reduces bias to zero while

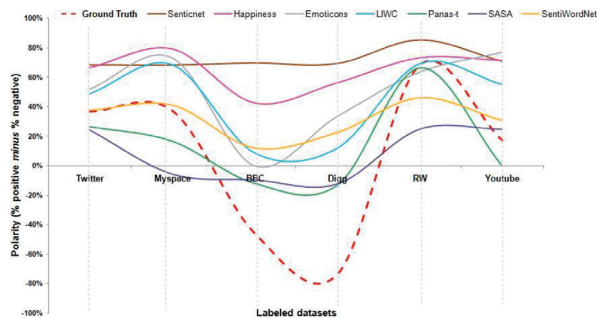


Figure 1: Polarity variation of seven sentiment analysis methods across different labeled datasets [6]

maintaining the prediction accuracy. Specifically, we make the following contributions in this paper:

1. We study the problem of prediction bias in lexicon-based polarity detection methods and propose a measure, named Polarity Bias Rate (*PBR*), for quantifying it.
2. We introduce BAT, an approach for easily controlling the bias in lexicon-based methods. This is the first method for addressing this problem in sentiment analysis.
3. We evaluate BAT on seven real-world datasets when combined with two popular lexicon-based methods, SentiStrength and AFINN. The results show that reducing bias often leads to improved prediction accuracy, thus further enhancing the utility of existing lexicon-based methods.
4. We demonstrate that the threshold can be learned reliably from a very small labeled dataset, while supervised methods learned on such small datasets perform poorly in comparison.

The rest of the paper is organized as follows. We discuss the related work in sentiment analysis and discrimination-aware data mining in Section 2. In Section 3, we propose a measure for quantifying polarity prediction bias and develop our approach for bias-aware sentiment analysis. Section 4 discusses the experimental evaluation of our approach, and Section 5 provides a concise discussion of our work. We conclude our paper in Section 6.

2. RELATED WORK

Sentiment analysis is the task of extracting and summarizing sentiments expressed in a document, while polarity detection or classification is the task of labeling a document as either positive or negative w.r.t. sentiment. Much work has been done on sentiment analysis and several methods have been developed for this purpose. Broadly, these methods can be divided into two types: supervised and lexicon-based. Supervised sentiment analysis methods require labeled data to learn a model that is then used to predict the sentiment for unseen documents. The major advantage of supervised methods is their ability to adapt to and learn from context given in the labeled data. Thus, labeled training data is essential for this type of methods but creating one

is usually tedious, time consuming, error prone, and/or prohibitively expensive for some domains. On the other hand, lexicon-based sentiment analysis methods use pre-compiled dictionaries of words with their intensity for positive or negative sentiment. These dictionaries are used to ascertain the sentiment of new documents.

Lexicon-based sentiment analysis methods, e.g., AFINN [11] and SentiStrength [16], have become very popular because of their unsupervised nature and easy-use properties. In the literature, different lexical dictionaries have been presented in different contexts, e.g., PANAS-t and POMS-ex [1] word lists were created for the Web context (informal writing) and LIWC [14] was created for formal English writing. Although lexicon-based methods do not require a labeled dataset for training, their coverage and performance can be affected by the context in which the word list is prepared and where it is used. Many lexicon-based methods have been proposed, but the literature suggests that SentiStrength [16] and AFINN [11] are the most popular methods.

Bias in sentiment analysis is a new research area but a substantial amount of work has been done in the related field of discrimination-aware data mining, first introduced in [13] [9]. Discrimination prevention, a key focus area in discrimination-aware data mining, studies techniques for making classifiers learned over biased/discriminatory datasets discrimination-aware. Similarly, in sentiment analysis we are interested in making polarity detection methods bias-free. As such, there are parallels to discrimination prevention techniques that involve classification algorithm tweaking [2] [9][8]. In [7], a decision theoretic framework is presented for making any classifier learned over biased datasets discrimination-aware at run-time. Although polarity bias has been identified in lexicon-based methods [6], no work has been reported on its analysis and control.

3. BIAS-AWARE SENTIMENT ANALYSIS

In this section, we develop our solution for bias-aware lexicon-based sentiment analysis. By bias in sentiment analysis, we mean the systematic preference for positive or negative polarity over the other in the predictions of a method. Such a bias is undesirable because it can mislead decision makers. We present an easy thresholding approach for correcting the bias without compromising on the benefits of lexicon-based methods. We start this section by describing briefly lexicon-based methods. Subsequently, we formally define polarity bias and present our solution for controlling bias in lexicon-based methods.

3.1 Lexicon-based Methods

Several lexicon-based sentiment analysis methods are available with different word lists and word valence values. However, only some recent methods contain words commonly used in user-generated content (e.g., slang, acronyms, etc). AFINN [11] and SentiStrength [16] are two such methods in popular use today. The latest version of the AFINN lexicon contains 2,477 unique words and phrases with valence values (sentiment strengths) ranging from -5 to $+5$, where -5 is the most negative and $+5$ is the most positive polarity strength. SentiStrength’s main lexicon contains 2,310 words and their stems. It also contains idioms list to modify strengths, spell correction list to correct spelling mistakes, repeated punctuation list to identify the emphasis (e.g., good!!!!!! is more positive than good because of

repeated exclamation marks), boosters word list (e.g., extremely poor is more negative than poor), negation list to identify negation of sentiment (e.g., I don't like you is negative), and emoticons list to detect sentiment expressed in symbols.

In general, consider a typical lexicon-based sentiment analysis method A . Such a method defines a list of words and phrases that convey polarity sentiment in documents. Each word and phrase in this sentiment lexicon or affective word list is given a valence value in the interval $[+v, -v]$ where v is a non-negative number that signifies the strength of polarity, and the sign of the valence value indicates the direction of the polarity (positive or negative). Such a lexicon of words/phrases and their valence values is usually developed manually by linguists.

Given a document d to be analyzed, method A generates two scores; a positive score, $S_A^+(d)$, and a negative score, $S_A^-(d)$. These scores are computed from the words and phrases in d that match those in the lexicon and by applying a combination operation on the valence values of these words/phrases (unmatched words/phrases are given zero valence values). Commonly used operations include maximum valence (e.g. SentiStrength) and weighted average valence (e.g. AFINN). Subsequently, document d is classified as either positive or negative by the following decision rule:

$$C_A(d) = \begin{cases} + & \text{when } S_A^+(d) > S_A^-(d) \\ - & \text{otherwise} \end{cases} \quad (1)$$

Lexicon-based methods are popular because they are unsupervised in nature (i.e., they do not require labeled training data).

Table 1: Truth table for polarity prediction

Actual↓, Predicted→	Pos (+)	Neg (-)
Pos (+)	TP	FN
Neg (-)	FP	TN

3.2 Bias in Sentiment Analysis

Table 1 shows the truth table for polarity predictions of the lexicon-based method A on $N = TP + FP + TN + FN$ examples. Each cell in this table gives the number of examples correctly or incorrectly predicted by method A . For example, TP (true positives) and FN (false negatives) are the numbers of examples correctly predicted as positive and incorrectly predicted as negative, respectively, by method A . Naturally, we prefer that method A makes as few errors as possible, i.e., $(FP + FN)$ is as small as possible (or accuracy is as high as possible). However, in many sentiment analysis applications, it is also critical that the number of false positives is not significantly greater than that of false negatives, or vice versa. Such a systematic prediction bias can distort the true sentiment makeup of the dataset.

Definition 1 (Polarity Bias Rate (PBR)): The Polarity Bias Rate (PBR) of a polarity classifier on N examples is defined as

$$PBR = \frac{FP - FN}{N}.$$

This value ranges from -1 to $+1$ where positive values indicate bias towards incorrect positive predictions.

PBR can also be expressed as the difference between the proportion of predicted positive examples $((TP + FP)/N)$

and the proportion of actual positive examples $((TP + FN)/N)$. For convenience, PBR can also be given as a percentage.

3.3 Our Solution: Bias-Aware Thresholding (BAT)

The goal of bias-aware lexicon-based sentiment analysis is to reduce PBR towards zero while maintaining high prediction accuracy (or low error rate). We present Bias-Aware Thresholding (BAT) for addressing this problem. BAT is motivated by cost-sensitive classification theory that allows specific errors to be reduced by incorporating a prediction threshold [4] [7]. In other words, changing the prediction threshold modifies the cost associated with making one type of error over the other, which is uniform in standard classification (see Eq. 1), thus forcing fewer errors of one type over the other type.

BAT can be combined with any lexicon-based sentiment analysis method to yield the following decision rule:

$$C_{BAT}(d) = \begin{cases} + & \text{when } (S_A^+(d) - S_A^-(d)) > t \\ - & \text{otherwise} \end{cases} \quad (2)$$

Here, $t \in \mathfrak{R}$ is a controlling threshold. A positive value of t will penalize positive polarity predictions (by increasing cost of false positive errors) while a negative value of t will penalize negative polarity predictions (by increasing the cost of false negative errors).

The value of threshold t can be found accurately and efficiently from a small set of labeled data. This is a simple root finding problem where we seek the t at which PBR is zero. It can be solved efficiently by any line search algorithm such as bisection method. As demonstrated in our experimental evaluation, even very small labeled datasets yield values of t that produce low PBR on test data, while supervised classification methods perform poorly in comparison when trained on such small datasets. BAT is an easy to implement addition to any lexicon-based sentiment analysis method (see Algorithm 1).

Algorithm 1: BAT - Bias-Aware Thresholding

- 1: **Required:** Method A , Documents D , Threshold t
 - 2: **Output:** Label (+ or -) for documents d in D
 - 3: **for all** $d \in D$ **do**
 - 4: **if** $(S_A^+(d) - S_A^-(d)) > t$ **then**
 - 5: $C(d) = +$
 - 6: **else**
 - 7: $C(d) = -$
 - 8: **end if**
 - 9: **end for**
-

4. EXPERIMENTAL EVALUATION

We evaluate BAT in combination with AFINN and SentiStrength on seven real-world datasets. We show the relationship between accuracy and PBR with change in threshold t , and compare BAT with naive Bayes classifier when a small amount of labeled data is used for finding the threshold t and for training the classifier. We start with briefly describing the evaluation datasets and setup.

4.1 Datasets and Setup

Table 2 summarizes the key characteristics of our evaluation datasets. In all, we conduct experiments on seven

Table 2: Datasets and their key characteristics

Data Source	# Docs	% Pos	% Neg
Movie Review	2,000	50%	50%
BBC	1,000	90.1%	9.9%
Digg	1,084	80.5	19.5%
Runner World	1,046	53.7%	46.3%
Twitter	4,242	68.4%	31.6%
YouTube	3,407	51.1%	48.9%
Large Movie Review	25,000	50%	50%

datasets that are described next.

Movie Review Dataset: The movie review dataset [12] contains user comments about movies available at Internet Movie Database (IMDB). To allow automatic processing only those comments were collected whose rating is in machine readable format. And, to avoid bias because of low number of prolific reviewers different filters were applied, e.g., a limit was applied to 20 reviews per user per sentiment category. Movie review dataset contains 1,000 positive and 1,000 negative reviews.

Web 2.0 Labeled Dataset: The Web 2.0 dataset is a human labeled dataset that was made available by the SentiStrength research group [15]. It includes wide range of messages, tweets, reviews, and comments from different sources. We use collections from: (1) BBC forum, (2) Digg, (3) Runners World forum, (4) Twitter, and (5) YouTube. Each document in the collection is labeled with positive and negative sentiment scores. For the assignment of binary polarity labels to each document we subtract positive score from negative score and if difference is greater than zero then positive polarity is assigned and negative polarity is assigned otherwise.

Large Movie Review Dataset: The large movie review dataset [10] is a large collection of movie reviews. It contains a total of 50,000 reviews with an equal percentage of positive and negative reviews. The dataset is divided into test and train sets of 25,000 reviews each. We performed experiments using the train dataset.

Experimental Setup: We use Python implementation of AFINN and Java implementation of SentiStrength available from the respective websites. While using these implementations no additional preprocessing of the text documents is done. For naive Bayes classifier, we use the implementation provided in RapidMiner. Standard text preprocessing of tokenization, stop word removal, and stemming is performed on the text documents.

4.2 Results

We conduct two categories of experiments. First, we evaluate the performance of BAT with changing values of its threshold t . We do this for both BAT combined with AFINN and BAT combined with SentiStrength. Second, we compare the performance of BAT/AFINN with naive Bayes classifier. In these experiments, the naive Bayes classifier and the threshold t of BAT is learned from varying sizes of labeled datasets and performance is measured over the respective remaining test datasets. For performance, we report both PBR and accuracy values. The baselines for our evaluations are standard AFINN and SentiStrength (for first category of experiments) and naive Bayes classifier (for second category

of experiments).

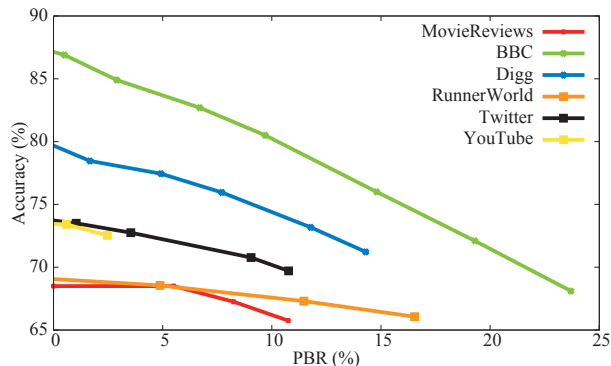


Figure 2: Performance of BAT when combined with AFINN

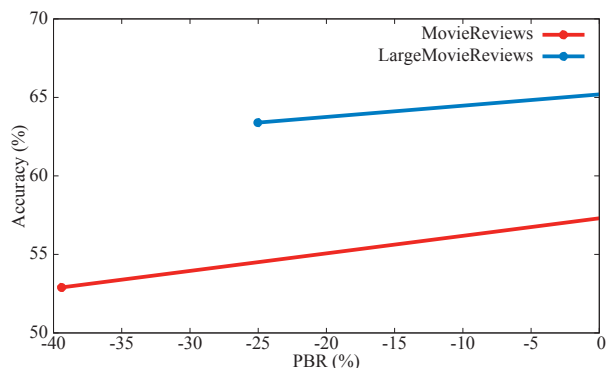


Figure 3: Performance of BAT when combined with SentiStrength

Figure 2 shows the performance of BAT when combined with lexicon-based method AFINN on different datasets. In this figure, prediction accuracy is given on the y-axis and prediction bias measured via PBR is given on the x-axis. Each line gives performances for a different dataset, and each point on the lines shows the accuracy and PBR values for a specific threshold t . The threshold $t = 0$ for the rightmost point on each line and t increases for points on the left. It is clear from this figure that standard AFINN (rightmost point on each line) exhibits strong bias towards positive sentiment. When AFINN is combined with BAT bias reduces gradually to zero and there is also a gradual increase in accuracy, with an increase in threshold t . This observation confirms that AFINN's predictions have a systematic bias and changing the prediction threshold not only reduces this bias but also increases prediction accuracy.

Figure 3 shows the performance of BAT when combined with SentiStrength. We do not report results of SentiStrength on the five Web 2.0 datasets because these datasets were used in the development of SentiStrength lexicon. Unlike the observation made for AFINN in Figure 2, SentiStrength exhibits a systematic bias towards negative sentiment (leftmost point on each line). When SentiStrength is combined with BAT and threshold t is decreased from zero bias reduces and accuracy increases. Again, this is a beneficial trend that is similar to that observed for BAT combined with AFINN.

The preceding experiments showed that different lexicon-based methods can have different biases on different datasets.

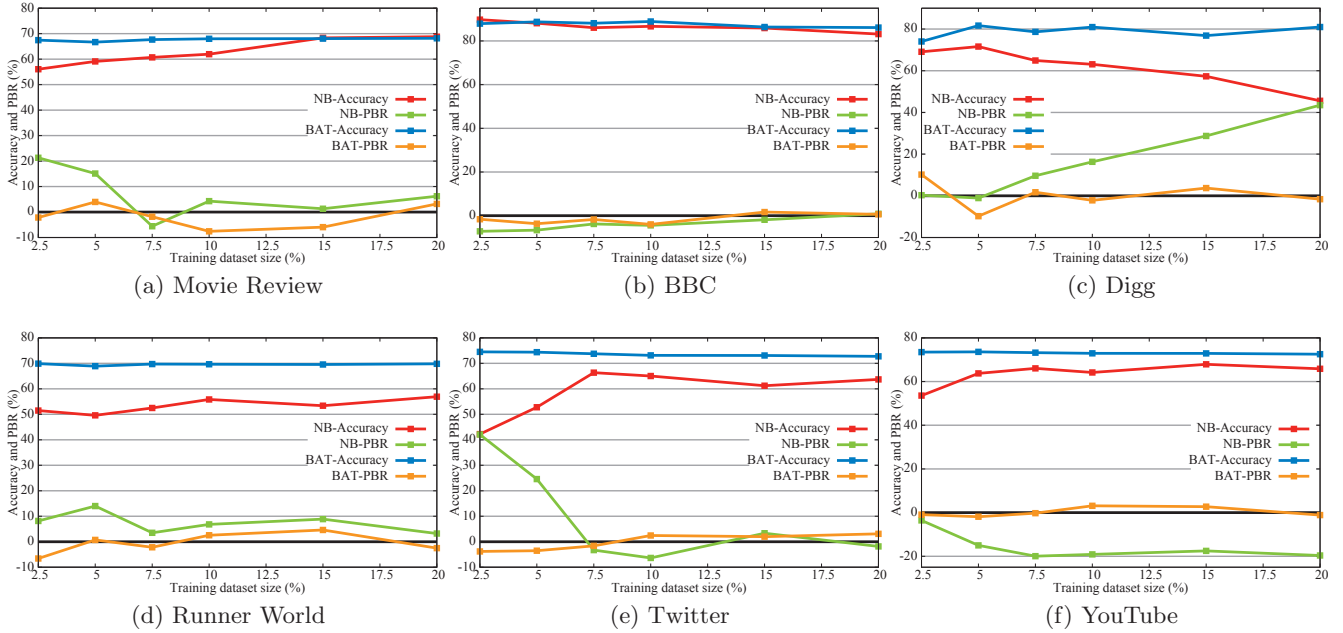


Figure 4: Comparison of BAT combined with AFINN and naive Bayes classifier using training dataset of varying sizes

To apply BAT combined with a lexicon-based method on a given dataset, it is necessary to find the threshold t that reduces the bias to a desired level (ideally to zero). For this purpose, some labeled examples for the dataset are needed. Once an appropriate threshold has been selected, this can then be used for predicting the polarity of new examples. Given this procedure, two questions arise: (1) how large should the labeled dataset be to reliably tune the threshold?, and (2) How well would a supervised method perform when learned on such a labeled dataset?

To address these questions, we conduct additional experiments. We hold-out varying sizes of each dataset and for each size we learn: (a) the threshold t that reduces the bias on this dataset to zero, and (b) a naive Bayes classifier on this dataset. We then evaluate the performances of BAT combined with AFINN using the learned threshold and the learned classifier on the remaining portion of the dataset.

Figure 4 shows the performance of BAT combined with AFINN and naive Bayes classifier over the test portions after learning from varying sizes of the datasets (each sub-figure shows results for one dataset). The x-axis in each sub-figure gives the training data size as a percentage of total data size and the y-axis gives the percent accuracy or PBR . The following observations can be made from Figure 4:

1. The threshold learned over the training data translates nicely to the test data by producing PBR close to zero.
2. Even when a very small training data size is used (2.5%) the performance of BAT combined with AFINN remains strong. In fact, there is no practically noticeable difference in PBR between 2.5% and 20% sizes of training data.
3. The naive Bayes classifier, on the other hand, produces varying non-zero biases. For some datasets, this bias is positive while for others it is negative.

4. More interestingly, in the vast majority of cases the accuracy of BAT combined with AFINN beats that for naive Bayes classifier.

5. DISCUSSION

Considering the importance of sentiment analysis in today's environment and the popularity of lexicon-based sentiment analysis methods these results hold much practical and theoretical significance. Firstly, it puts a question mark on some of these sentiment analysis methods. Perhaps their scoring and aggregating strategy needs to be revisited so as to reduce bias, or perhaps they should always be paired with some supervision as is obvious from our approach. This also demands more extensive evaluation of sentiment analysis methods. Secondly, it is somewhat surprising to see naive Bayes classifier, which is a supervised method, also producing significant bias and lower accuracy. This suggests that lexicon-based methods do have much potential and should be explored further by researchers.

Finally, from a practical perspective BAT when combined with a lexicon-based method like AFINN gives reliable performances with respect to both accuracy and bias. This is a great advantage especially considering that labeling a few dozen examples for training produces reliable estimates for the threshold that reduces bias towards zero and improves accuracy.

6. CONCLUSION

Systematic bias in polarity predictions can jeopardize decisions based on automatic sentiment analysis of textual content. Biased polarity prediction methods can produce excessive false positives or false negatives that distort the true sentiment distribution of the dataset. In a recent paper, it was been shown that lexicon-based sentiment analysis methods, which are very popular today because of their ef-

iciency and effectiveness, can produce biased predictions. However, till now no work has been reported on addressing this problem in such methods.

In this paper, we study the problem of bias in lexicon-based polarity prediction in detail and propose an approach for its solution. We define a measure, called Polarity Bias Rate, for quantifying the bias. Our approach, called Bias-Aware Thresholding (BAT), combines with any lexicon-based method to make it bias-aware. Specifically, BAT introduces a prediction threshold that penalizes systematic errors to reduce bias and improve accuracy. Our approach is simple yet effective, and can be readily used in practice. We evaluate our approach on seven real-world datasets when combined with AFINN and SentiStrength. Our extensive empirical evaluation confirms that our approach controls bias effectively while maintaining (usually improving) prediction accuracy.

This topic needs greater attention from researchers. The reasons for biases and more effective solutions for their control need to be investigated. Moreover, extensive experimental evaluations and their relation to linguistics may yield additional insights into the problem.

References

- [1] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, 2009.
- [2] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.
- [3] E. Cambria, R. Speer, C. Havasi, and A. Hussain. Senticnet: A publicly available semantic resource for opinion mining. *AAAI Fall Symposium: Commonsense Knowledge*, 10(0), 2010.
- [4] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.
- [5] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining.
- [6] P. Gonçalves, M. Araujo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. *Proceedings of the First ACM Conference on Online Social Networks*, (12):27–38, 2013.
- [7] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. *12th IEEE International Conference on Data Mining, ICDM 2012*, pages 924–929, December 2012.
- [8] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. *Proceedings of The 3rd IEEE International Workshop on Privacy Aspects of Data Mining*, pages 643–650, 2011.
- [9] B. T. Luong, S. Ruggieri, and F. Turini. k-*nn* as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 502–510, 2011.
- [10] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, June 2011.
- [11] F. Å. Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Micro-posts': Big things come in small packages*.
- [12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10(8):79–86, 2002.
- [13] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, 2008.
- [14] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, March 2010.
- [15] M. Thelwall. Heart and soul:sentiment strength detection in the social web with sentistrength, <http://sentistrength.wlv.ac.uk/documentation/sentistrengthchapter.pdf> (remove hyphen).
- [16] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December 2010.
- [17] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, (8):417–424, 2002.
- [18] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, 2012.