A Comparison of Dictionary Building Methods for Sentiment Analysis in Software Engineering Text

Md Rakibul Islam
University of New Orleans, USA
Email: mislam3@uno.edu

Minhaz F. Zibran
University of New Orleans, USA
Email: zibran@cs.uno.edu

***Abstract*—Sentiment Analysis (SA) in Software Engineering (SE) texts suffers from low accuracies primarily due to the lack of an effective dictionary. The use of a *domain-specific* dictionary can improve the accuracy of SA in a particular domain. Building a domain dictionary is not a trivial task. The performance of lexical SA also varies based on the method applied to develop the dictionary. This paper includes a quantitative comparison of four dictionaries representing distinct dictionary building methods to identify which methods have higher/lower potential to perform well in constructing a domain dictionary for SA in SE texts.**

I. Introduction

Software engineering (SE) studies [4], [5], [6] involving sentiment analysis (SA) in text repeatedly report concerns about the accuracy of those SA tools. The widely used SA tools (e.g., SentiStrength, NLTK) rely on a dictionary-based lexical approach, where sentiment lexicons constitute the core elements of the dictionary. The low accuracies of those SA tools are largely due to the limitations of the underlying dictionary [6] especially when operated in a technical domain such as SE. The limitations of the dictionary mainly stem from the domain specific variations in the meanings of technical words in informal SE text [6]. We must build domain-specific dictionaries to improve the accuracy of SA of text in a particular domain [6].

To build a dictionary, several approaches [1], [2], [7], [14] have been attempted in the past resulting in variations in their performances [7], [11], [14]. The construction of a domain dictionary is a tedious task [14] and it is often difficult to know in advance which approach for dictionary building can suite better for a particular domain [2]. A few studies were conducted on texts from non-technical domains (e.g., movie and restaurant reviews) to find out appropriate dictionaries for those domains [7], [10]. However, no such study was performed for a technical domain such as SE.

This paper presents a comparative study of four general-purpose dictionaries representing four distinct methods from dictionary creation. In particular, we examine if these dictionaries, as created using different methods, show substantially different results in sentiment detection in SE texts. The results from the study will help in identifying potentially suitable methods for the creation of a domain dictionary for SE texts.

II. Methodology

In this study, we use our recently developed SA tool `SentiStrength-SE` [6]. This lexical tool is developed in a modular manner to allow replacing its default dictionary with a different one. `SentiStrength-SE` is reported to have substantially higher accuracy in SA in SE text compared to the most widely adopted tool in SE community [6]. This is one of the reasons why we choose this tool for our study.

Dictionaries under Study: In this work, we study four different dictionaries (`SentiStrength` [12], `AFINN` [8], `MPQA` [13], and `VADER` [3]). We select four dictionaries based on their well-establishment in SA and recency of their development. In addition, we make sure that those selected dictionaries are developed using distinct methodologies.

Although all the selected dictionaries are developed by leveraging previously created dictionaries, still those differ in their construction methodologies and contents. `AFINN` [12] starts with a collection of seed words and expands that by including appropriate words from microblogs and other sources [8]. `VADER` [3] starts with collecting words by examining microblogs and existing dictionaries LIWC, GI and ANEW [11]. The reliability of lexicons in `VADER` are assessed using 10 human raters, whereas no such reliability assessment by human is done for the `AFINN` dictionary. The `SentiStrength` [12] dictionary is constructed by combining LIWC and GI dictionaries similar to `VADER`, and also includes lists of emoticons, negations and intensifiers.

`MPQA` [13] uses GI dictionary as a basis and expands including more sentimental words. In `MPQA`, the polarities of words are determined based on contexts using parts-of-speech (POS) and subjectivity of words, whereas `SentiStrength`, `VADER` and `AFINN` dictionaries are developed without inclusion of such contextual sense.

Benchmark Dataset: We use a benchmark dataset [9], which is the only publicly available such dataset in the SE domain [6], [9]. In our work, we use the Group-2 and Group-3 portions of the dataset [9] containing 1,600 and 4,000 issue comments extracted from JIRA issue tracking system. The issue comments are manually annotated with six basic emotions (i.e., joy, love anger, sadness, fear, and surprise), which we translated to sentimental polarities (i.e., positive or negative). Further elaboration about the translation to sentimental polarities can be found in our earlier work [6] and details about the benchmark dataset can be found elsewhere [9].

Study Procedure: For each of the four dictionaries, we configure `SentiStrength-SE` to use that particular dictionary, and run the tool on issue comments. The outputs of the tool are compared with the annotated benchmark dataset to compute

TABLE I
COMPARISON AMONG THE FOUR DICTIONARIES UNDER STUDY

Data	Sentiment	Metric	SentiSt.	AFINN	MPQA	VADER
Group-2	Positive	Precision	74.48%	**79.49%**	65.47%	64.27%
		Recall	98.81%	**99.08%**	98.68%	**99.08%**
		F-score	84.93%	**88.21%**	78.72%	77.97%
	Negative	Precision	28.22%	19.91%	22.60%	**30.49%**
		Recall	97.66%	35.94%	**99.22%**	97.66%
		F-score	43.78%	25.63%	36.81%	**46.47%**
	Neutral	Precision	96.83%	88.57%	**97.57%**	97.00%
		Recall	52.42%	**66.43%**	40.14%	37.00%
		F-score	68.01%	**75.92%**	56.88%	53.57%
Group-3	Positive	Precision	**31.69%**	30.51%	18.39%	19.77%
		Recall	87.79%	82.04%	85.52%	**88.20%**
		F-score	**46.58%**	44.48%	30.27%	32.30%
	Negative	Precision	47.61%	**48.69%**	36.61%	47.97%
		Recall	78.40%	50.45%	76.59%	**78.70%**
		F-score	59.25%	49.55%	49.54%	**59.61%**
	Neutral	Precision	91.28%	85.29%	90.97%	**91.44%**
		Recall	56.16%	**68.74%**	41.62%	48.12%
		F-score	69.54%	**76.13%**	57.11%	63.06%
Overall average accuracy		Precision	**61.69%**	58.74%	55.27%	58.49%
		Recall	**78.54%**	67.11%	73.63%	74.79%
		F-score	**62.02%**	60.00%	51.55%	55.50%

the performance of each dictionary in terms of the precision, recall (and F-score) in the detection of positive, negative, and neutral sentiments.

III. RESULTS

Table I presents the precision, recall and F-score of sentiment detection at separate run of SentiStrength-SE using each of the four dictionaries. Looking at the results, it is not easy to distinguish a clear winner. Considering the overall average accuracy, as presented at the bottom of the table, the SentiStrength dictionary appears to have performed the best, followed by AFINN and VADER. Notice that, although AFINN performs slightly better than SentiStrength in detecting positive and neutral sentiments, its performance is much worse compared to SentiStrength in detection of negative sentiments. Thus, AFINN falls behind the SentiStrength dictionary in overall accuracy.

It is interesting that MPQA is found to have performed the worst, although the dictionary incorporates contextual information and subjectivity. SE text being informal, containing technical jargons, often including misspelled words and grammatically incorrect sentences can be among the reasons for MPQA's poor performance in our study. Thus, it appears that simple lexicon-based approaches for dictionary creation work better for SA in SE text.

IV. THREATS TO VALIDITY

The dictionaries included in our study are all general-purpose, meant to work for text in any domain *including SE*. Although, their comparison indicate which method of dictionary creation might work better for SE text, a higher construct validity could be achieved if those were domain dictionaries specific for SE. This was not possible because only one SE-specific domain dictionary is known to exist [6].

The results are derived based on an analysis over only one dataset. This can be a threat to the external validity of the study. All the dictionaries, tool, and issue comments used in this work, are publicly available. Thus, we develop high confidence in the reliability of the study.

V. RELATED WORK

Several work [1], [2], [14] in the past created new dictionaries and compared against existing ones. Kim et al. [7] compared four domain-specific dictionary construction methods using movie review data. Regan et al. [11] quantitatively measured the accuracies of six dictionary-based methods by applying to four different corpora. Pablos et al. [10] also compared the performances of 10 dictionaries, which were developed using different methodologies. While all the mentioned studies compared dictionary development methodologies using non-technical data, for the first time, we have done such using text from a technical domain.

VI. CONCLUSION

We have studied the prospect and effectiveness of four distinct dictionary building methods for sentiment detection in SE texts. Based on a quantitative analysis over a benchmark dataset, we have found that dictionaries (i.e., SentiStrength, AFINN, and VADER) created using simple lexicon-based approach perform better (for SA in SE text) than those (i.e., MPQA), which include complex techniques for incorporating subjectivity and contextual sense.

Since our study is based on only one fair-size dataset, further larger scale studies can be conducted to verify or validate the results. This remains within our immediate future work. We also plan to create a large sentiments annotated dataset of SE texts, which will enable carrying out similar studies in a larger scale.

REFERENCES

[1] J. Bross and H. Ehrig. Automatic construction of domain and aspect specific sentiment lexicons for customer review mining. In *CIKM*, pages 1077–1086, 2013.

[2] H. Hammer, A. Yazidi, A. Bai, and P. Engelstad. Building domain specific sentiment lexicons combining information from many sentiment lexicons and a domain specific corpus. In *CIIA*, pages 205–216, 2015.

[3] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *WSM*, pages 216–225, 2014.

[4] M. Islam and M. Zibran. Exploration and exploitation of developers' sentimental variations in software engineering. *Internation Journal of Software Innovation*, 4(4):35–55, 2016.

[5] M. Islam and M. Zibran. Towards understanding and exploiting developers' emotional variations in software engineering. In *SERA*, pages 185–192, 2016.

[6] M. Islam and M. Zibran. Leveraging automated sentiment analysis in software engineering. In *MSR*, pages 203–214, 2017.

[7] M. Kim, J. Kim, and C. Juing. Performance evaluation of domain-specific sentiment dictionary construction methods for opinion mining. *Intl. J. of Database Theory and Application*, 9(8):257–268, 2016.

[8] F. Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *MSM*, 2011.

[9] M. Ortu, A. Murgia, G. Destefanis, P. Tourani, R. Tonelli, M. Marchesi, and B. Adams. The emotional side of software developers in JIRA. In *MSR*, pages 480–483, 2016.

[10] A. Pablos, M. Cuadros, and G. Rigau. A comparison of domain-based word polarity estimation using different word embeddings. In *LREC*, pages 54–60, 2016.

[11] A. Reagan, B. Tivnan, J. J. Williams, C. Danforth, and P. Dodds. Benchmarking senti. anal. methods for large-scale texts: A case for using continuum-scored words and word shift graphs. *ArXiv e-prints*, 2015.

[12] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Info. Science and Tech.*, 63(1):163–173, 2012.

[13] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level senti. analysis. In *HLT/EMNLP*, pages 347–354, 2005.

[14] L. Young and S. Soroka. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29:205–231, 2012.