A Domain-Dependent Lexicon to Augment CAD Peer Review

Zachariah J. Beasley[1] and Les A. Piegl[2]

[1]University of South Florida, zjb@mail.usf.edu
[2]University of South Florida, lpiegl@gmail.com

Corresponding author: Les A. Piegl, lpiegl@gmail.com

Abstract. CAD course creative projects necessitate subjective feedback. In academia, peer review is a widely used instrument to gather diverse and timely feedback which stimulates learning and engagement in students who review one another. To date, however, no effort to summarize and score subjective content from peer review text via sentiment analysis has been attempted in an educational setting, including CAD courses — many of which naturally employ a project-based architecture. This is perhaps due in part to a lack of specifically tuned tools. Towards meeting this need, we introduce a new lexicon compiled from actual peer review text, implemented specifically in a CAD-course context, and compare it to other publicly available lexicons. HeLPS, our domain-dependent lexicon, performed more concisely and accurately in our CAD courses and consistently tagged high-quality positive and negative sentiment with a lexicon a fraction of the size of others. Both qualitative and quantitative evidence suggest that HeLPS is the preferred option for identifying subjective opinion towards CAD course projects.

Keywords: sentiment analysis, CAD education, dictionary, peer review, educational data mining
DOI: https://doi.org/10.14733/cadaps.2021.186-198

1 INTRODUCTION

Grading creative works requires subjective analysis. Since creative works are often intended for a wider audience than a single instructor, it is beneficial to gather an audience (crowd) response. In academia, peer review is a widely used mechanism to gather diverse and timely feedback which stimulates learning and engagement in reviewing students [3], [22], [28]. While peer review numerical data is valuable, students' open-ended comments can contain a wealth of information not captured in the review form. To date, however, no effort to summarize and score subjective content from peer review *text* via sentiment analysis has been attempted in an educational setting, including CAD courses — many of which lend themselves easily to a project-based architecture. This is perhaps due in part to a lack of specifically tuned tools. Leveraging information from

peer review text is an open challenge that, if solved, can assist with grading in CAD courses by providing more information to an instructor [4], [33].

Any system that attempts to understand text must balance an inverse relationship between breadth of coverage and depth of understanding. This dichotomy is analogous to a choice between recall and precision in machine learning. When deciding between a lexicon (dictionary) or neural network approach for sentiment analysis [40] found that "dictionaries had exceptional precision, but very low recall, suggesting that the method can be accurate, but that current lexicons are lacking scope. Machine learning systems worked in the opposite manner, exhibiting greater coverage but more error." Our emphasis on precision motivated the choice of a lexicon-based approach in addition to its intuitiveness, interpretability (since the rationale for a grade was occasionally requested by students), and accuracy on short segments of text. See [6] for some other limitations of neural networks, especially in identifying linguistic properties and explaining predictions.

In building our domain-dependent lexicon to find subjective content, we prioritized depth of insight and precision over breadth and recall since our system produced a grade (i.e. high stakes). While we required a certain level of coverage to ensure confidence in assigning a comment grade per review, we intentionally excluded over-used words that served as noise rather than providing quality information. Even within highly related courses (e.g., Computer Graphics and Software Testing, both in the Computer Science domain), we found it necessary to modify our lexicon to maintain contextual polarity [48]. For example, in Software Testing *bug* and *fix* were often not negative words as they are in general vernacular. Instead, they typically indicated that a team had successfully found and corrected a seeded fault (i.e. a positive sentiment). Thus, we agree with the general sentiment of previous work (e.g., [17]) that the process, rather than the lexicon itself, should be copied. Although this creates overhead per course, some subjects are similar enough to share a reduced lexicon from the intersection of important words in two different courses. The effort to construct a domain-dependent lexicon is further mitigated through the use of an aspect extractor detailed in a prior work [3].

The main contributions of this paper are 1) a brief overview of a process to create a weighted domain-dependent lexicon from student peer review text, implemented specifically in a CAD-course context and 2) an evaluation of the lexicon against others publicly available. This work is relevant to any course which gathers student input from peer review, regardless of whether the course is project-based or peer review is used to provide a grade, because it details a process for increasing the amount and accuracy of information gained from written text. The structure of the courses [33], design of the review form ([5], [33]), and implementation of the sentiment analysis algorithm [5] are covered more fully in prior work and thus only briefly mentioned when relevant.

In section 2, we cover related work: applications gathering sentiment within academia as well as the potential for taking a trained sentiment analysis algorithm from another domain and using it in an academic environment (transfer learning). Section 3 outlines the process by which we built our domain-dependent lexicon, HeLPS: the Heuristic Lexicon of Peer Sentiment. In section 4, we describe two general formatting classes for publicly available lexicons and describe in detail those we compared against. Section 5 presents the metrics used for evaluation and three quantitative experiments with a qualitative example to determine the usefulness and precision of our lexicon in contrast to others. Finally, in section 6 we conclude with our findings.

2 RELATED WORK

Sentiment analysis is no stranger to the classroom. Recently, there have been a number of interesting applications:

- Student predicted attrition in Massively Open Online Courses (scored by a product reviews lexicon [46] or SentiWordNet 3.0 [10])
- Student subjective perception of a teacher (scored via Naïve Bayes [14] or an ensemble [34])

- Students who have issues with a course (scored by the Microsoft Text Analytics API [37] or a mixed graph of terms [11])
- Student identification of teacher weaknesses and strengths (scored by Naïve Bayes [2] or a lexicon in R [35] or an ensemble [21])
- Student internship experience evaluation (scored manually [16])

While the previous applications are all potentially valuable for gathering information on a teacher or student, none contribute to summarizing and scoring subjective peer review content for a grade, increasing the reliability of the grading process, or providing a sentiment lexicon for classroom use.

A number of these applications rely on lexicons or algorithms trained in other domains, since sentiment analysis is widely used in social media, creative works (e.g., movies), and products. In our context, the stakes are higher when generating a student grade so there are both ethical and practical considerations when taking a trained classifier from an outside domain and using it to generate student grades from subjective content.

Ethically, grades based on a trained classifier or lexicon from another domain seems unfair. This concern is heightened when authors note that sentiment analysis is "highly sensitive to the domain from which the training data is extracted" [24] and that a domain-dependent lexicon is necessary to ensure accuracy does not "suffer a serious performance loss once the domain boundary is crossed" [25]. Others have found wide variance when applying the same industry-standard sentiment analysis tools to different domains (e.g., an accuracy of 80.4%, 71.5%, and 62.5% on movie reviews, Amazon reviews, and tweets, respectively [38]) or even within domain using a hybrid (local and general-purpose) lexicon [26]. Finally, other research suggests that "methods are often better in the datasets [in which] they were originally evaluated" even for popular algorithms like SentiStrength ([43], [44]) and SO-CAL [42].

In a comprehensive paper covering twenty-four algorithms and eighteen datasets (e.g., Youtube, Twitter, Yelp, Myspace, and Amazon) to determine sentiment analysis benchmarks, it was found that "sentiment analysis methods cannot be used as 'off-the-shelf' methods, especially for novel datasets" [36]. Specifically, when narrowed down to the software engineering domain (e.g., Jira and Stack Overflow), some researchers claim that generic sentiment analysis tools are inadequate and exhibit disagreement while domain-specific tools increase accuracy [30]. Finally, a few authors simply label the entirety of state-of-the-art sentiment analysis tools deficient in the software engineering domain [23].

Others, in contrast, found significant improvement with their crowd-labeled, domain-specific negative word lexicon compared to two other general-purpose lexicons classifying negativity in German media reports and party statements [17]. Similarly to us, they advocate a process and note that even "some commercial providers advise against using their sentiment lexicon out-of-the-box without customizing it to the domain" [17]. The sum of this research connotes a "clear benefit to creating hand-ranked, fine-grained, multiple-part-of-speech dictionaries for lexicon-based sentiment analysis" [41] and underscores the necessity of a domain-dependent lexicon for accuracy in educational sentiment analysis.

3 BUILDING THE LEXICON

Our mixed graduate and undergraduate project-based CAD courses were organized to increase the number of peer reviewers per group presentation, essay, or term project (typically 25-35) [33]. This provided a wealth of information from which to draw key words and concepts for our domain-dependent lexicon. More information on the course design and data-driven development of the review form can be found in a previous work [5].

3.1 Key Word Selection

Over six semesters (eleven courses, both CAD and non-CAD with approximately 425 students), we gathered sentiment-bearing key words by hand from the reviews to add to our lexicon, HeLPS: the Heuristic Lexicon of

Positive	Negative	Negate	Flag
superb	clumsy	miss	copying
fascinate	erroneous	not	cheated
innovative	superficial	wish	cheater
accurate	omission	hardly	plagiarism
nicely	mistake	suggest	plagiarize

Table 1: Sample of Lexicon Words

Peer Sentiment. In contrast to the review form questions, which were selected for their breadth, we selected any relevant words — even those used very infrequently — to add to our lexicon ([32] also found a balance of frequent and rare words necessary to discover subjective content). Human intelligence was required for this task — we could not simply select the most common feedback (e.g., *good* or *bad*) because it did not add meaningful information. Instead, we intentionally cut through the noise and selected only words that provided rich meaning. This was done through the process of intelligent data combing: *selecting information-rich key words and phrases, through human intelligence, to correctly analyze and summarize student observations* (similar to the *concept-based approach* leveraging a human knowledge base in [9]). This process was later augmented and semi-automated through *aspect extraction* (similar to [7] and described in [3]). Table 1 shows some sample key words from our lexicon.

Key words were then stemmed and weighted by instructor heuristic (see subsection 3.2). Thus, HeLPS could be interpreted as a stemmed seed set which was not expanded through a lexical learning strategy since we desired a smaller set of words specific to our domain and weighted by heuristic (in contrast to [18], [19], and [45]). We currently have 283 positive and 207 negative words in our lexicon. We have an additional 22 words that negate sentiment and 11 that are flag words like *cheating*. The inclusion of flag words allows us to crowdsource plagiarism detection. These word sets are much smaller than the smallest out-of-domain lexicon we compared in section 4. However, because the words were chosen in-domain, the information generated compared very favorably to other, larger lexicons (section 5).

We tracked the variety of key words used per group presentation, essay, or term project and per semester, as well as the percentage of lexicon matched per student work (typically 18-20% of positive words and 4-8% of negative words). Figure 1 shows a word cloud from WordArt.com of a full semester of mixed positive and negative key words from our CAD Modeling course. Word size correlates to the number of mentions. The top key word students wrote was *example*, which was recorded 6,141 times. The least-mentioned key words were *dull* (mentioned twice), *regurgitate* (mentioned twice), and *eliminate* (mentioned once). Figure 1 exemplifies both general (e.g., *understand* and *useful*) as well as domain-specific key words (e.g., *cite* and *diagram*).

3.2 Polarity

The polarity of a word, phrase, or sentence is comprised of direction (positive, negative, or neutral) and an optional weight.

3.2.1 Direction

To classify the direction of key words found during intelligent data combing, we grouped tokens (words and punctuation) into six sets based on context within the sentence: positive word, negative word, neutral word (not stored in any dictionary), negate word, flag word, and reset token. Figure 2 demonstrates how each set fit into three sentiment directions — positive, negative, neutral. While reset tokens were solely neutral (e.g.,

Figure 1: Word Cloud of Semester Key Words

but), negative and negate words overlapped (e.g., *lacked*) and so did negative and flag words (e.g., *cheating*). Our negation strategy used intuitive rules (both local and global, similarly to [27], [31], [47], and [48]) and was applied uniformly to every lexicon against which we compared.

3.2.2 Weight

After the direction of sentiment was determined, key words were weighted by instructor heuristic as opposed to a learning strategy like [1] and [49]. Other lexicons are similarly weighted by an expert [29], small group [8], or a crowd (e.g., Amazon Mechanical Turk workers in [17], [20], and [39]). For simplicity, both positive and negative key words were weighted on a continuous scale between zero and one.

4 OTHER LEXICONS

A variety of general lexicons for sentiment analysis currently exist. A handful of lexicon/software combinations are proprietary and for licensed use only (e.g., WordStat and Linguistic Inquiry and Word Count (LIWC)) but many are available for academic purposes. The two primary formats for lexicons are as follows:

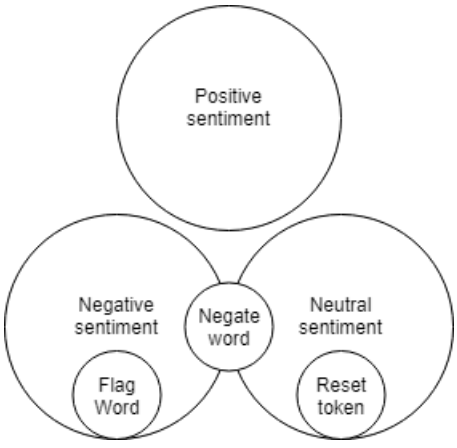

Figure 2: Polarity of Tokens

4.1 Synsets

A synset is a semantically-related set of words — *cognitive synonyms* — that share meaning. A synset is comprised of an ID that uniquely identifies the row as well as a positive value, a negative value, (potentially) an objectivity value (1 - (Pos-value + Neg-value)), a *gloss* which includes one or more words that share a meaning, a definition, and one or more examples. The fundamental thought is that polarity can be found by examining related words — positive words are connected to other positive words and vice-versa. An example synset is as follows:

{02084101 0.5 0.25 studious#2 bookish#1 characterized by diligent study and fondness for reading; "a bookish farmer who always had a book in his pocket"; "a quiet studious child"}

SentiWordNet 3.0 [1] and Micro-WNOp are examples of synset lexicons that provide sentiment scores to synset entries from Princeton's WordNet, an unlabeled lexicon.

4.2 Word-Value Pair

In contrast to synsets, word-value pair lexicons do not store information about other semantically-related words and assume that a word has either an exclusively positive or negative value. This simplifies the calculation of sentiment and negation, although it can be argued that it does forgo some contextual elasticity since a particular word can be used in different ways. To limit the impact of choosing an exclusive sentiment direction, lexicons are often domain-specific. Examples of such lexicons include Affective Norms for English Words (ANEW) [8], SlangSD [49], Multi-Perspective Question Answering (MPQA) [13], Valence Aware Dictionary and sEntiment Reasoner (Vader) [20], and AFINN-111 [29]. Interestingly, SlangSD incorporates slang words scraped from UrbanDictionary.com and Vader includes emoticons (e.g., :D), word shape (e.g., all caps to denote passion), punctuation that increases intensity (e.g., !!!), and acronyms (e.g., lol). For an academic domain such as student peer reviews we have not noticed situations in which the additional breadth of these lexicons is necessary, however. Finally, MPQA includes misspelled words that the authors found frequently occurred.

Due to their simplicity, domain-centric focus, and intuitive design we chose to implement HeLPS as the latter class of lexicons, the word-value pair. Any lexicons in synset format were converted to word-value pair format by extracting only words with an exclusively positive or negative direction.

4.3 Formatting

Each lexicon had slight differences that required adaptation to match our format. For every lexicon that included a numeric value for sentiment, we produced both a weighted and unweighted variant. When mapping from a different scale to our scale of [0, 1] we mapped [0.1, 1] so that the least-weighted positive and negative words would not receive a weight of 0. Finally, none of the other lexicons were stemmed, thus review comments were left un-stemmed for comparison. The descriptions and formatting of each lexicon we tested are as follows:

- AFINN-111: manually labeled by Finn Arup Nielsen in 2009-2011. Words were originally scored on a scale of [-5, 5]. Only words with a sentiment [-5, -2] and [2, 5] were kept. The resulting tagged word set was thus reduced from 2,477 to 670 positive and 1,289 negative words.
- ANEW-2017: manually labeled by a group of introductory psychology students for class credit. Words were originally scored on a scale of [0, 9]. Only words with a sentiment [0, 3] and [6, 9] were kept. The resulting tagged word set was thus reduced from 3,189 to 1,162 positive and 436 negative words.
- MPQA: automatically labeled by using a seed set with provided polarity orientations expanded through a WordNet search for semantically related words. Unlike the other lexicons, words did not receive a fine-grained score. Thus, the word sets did not have to be mapped numerically and only an unweighted lexicon was extracted. The number of tagged words remained the same at 2,007 positive and 4,783 negative words.
- SentiWordNet 3.0: automatically labeled similarly to MPQA, using a semi-supervised approach with WordNet. Words were originally scored on a scale of [0, 1] and thus were not scaled. Only words with an exclusive sentiment (positive or negative) in the range of [0.1, 1] were kept. The number of tagged words was thus reduced from 117,659 to 9,069 positive and 9,601 negative words.
- SlangSD: automatically labeled 1) via existing lexicons — SentiWordNet, LIWC, MPQA, and a previous work (1% of the lexicon) 2) leveraging Twitter (average of the nearest words in 150 tweets containing the target word, 23% of the lexicon), and 3) from a seed set of related words expanded on UrbanDictionary (76%). Words were originally scored on a scale of [-2, 2]. Only words with a sentiment [-2, -1] and [1, 2] and of length one were kept. The resulting tagged word set was thus reduced from 96,461 to 10,479 positive and 29,025 negative words.
- Vader: manually labeled by ten independent raters. Words were originally scored on a scale of [-4, 4]. Only words with a sentiment [-4, -2] and [2, 4] were kept. The resulting tagged word set was thus reduced from 7,517 to 1,009 positive and 1,237 negative words.

It is interesting to note that, with the exception of HeLPS and ANEW, negative lexicon length surpassed that of the corresponding positive lexicon.

5 EXPERIMENT

We compared the aggregation of sentiment between our lexicon and six others publicly available — AFINN-111 [29], ANEW-2017 [8], MPQA [13], SentiWordNet 3.0 [1], SlangSD [49], and Vader [20] — by holding the scoring algorithm constant. In brief, each review text score was simply the sum of sentiment words, considering negation, scaled to a common range (see [5] and [33] for details and [47] for a similar negation strategy). Each project's textual sentiment score was the mean of all review sentiment scores.

We assumed that the aggregate sentiment score would be in the general proximity of the aggregate review form analytical score. Thus, we primarily evaluated the mean absolute error between the aggregate form score and the mean or median sentiment score. Note that we did not compare a single student's review comment sentiment score to their corresponding review form analytical score, rather, the aggregation of each

Table 2: Weighted (W) vs. Unweighted (UW) Lexicon

Course	LowerStddev	LowerMAE	LowerMdAE	HigherAM
CG	W	W	W	W
CM	W	W	W	W

from 25-35 reviews. We found students used the analytic and subjective parts of the review form differently (a phenomenon also observed by [12] in the product reviews domain), precluding us from using the review form score as a ground truth per comment. Others have also discovered this discrepancy when attempting to match the text sentiment to a ground-truth label a customer actually chose [15]. To check lexicon accuracy, we contrast the following attributes on our two most recently completed mixed graduate and undergraduate CAD courses, Computer Graphics (CG, 36 student) and CAD Modeling (CM, 31 students):

- FormScore (FS) is the mean of a collection of reviewers' 3-option radio button responses towards a single student work: [0, 4.3]
- Mean/MedianSentiment (MS/MdS) is the mean/median of a collection of review comment sentiment scores towards a single student work: [0, 4.3]
- AvgMatch (AM) is the average percentage of student works (36 per semester) where FS has the same letter grade as MS: [0, 1]
- Mean/MedianAbsError (MAE/MdAE) is the absolute difference between FS and MS/MdS, an accumulation of error in works over a given course: [0, inf)

5.1 Weighted vs. Unweighted

Intuitively, an intelligently-weighted lexicon should outperform an unweighted one. Additionally, a lexicon weighted by an instructor should more closely match their desire when processing text. Table 2 demonstrates a comparison between some key positive attributes when using the weighted vs. unweighted lexicons: lower standard deviation, lower MAE, lower MdAE, and higher AM. For each metric, the weighted versions were found superior.

Standard deviation fell by 55-56%, mean absolute error was reduced by 57-67%, and median absolute error lowered 45-49%. Although not an indicator of correctness, we noticed the unweighted versions were slightly more generous, with a higher average mean (3-4%) and median (5%) than the weighted versions.

5.2 Matching Form Score

Because the weighted versions were found to have superior general qualities, we used them to compare between lexicons. Table 3 shows the lexicon mean/median absolute error and average matched, with the best scores bolded. In both courses, HeLPS was first in a majority of metrics. Most importantly, we found HeLPS had the lowest mean absolute error. ANEW (also weighted by human heuristic) appeared to be the next most accurate lexicon in the majority of cases. The two largest lexicons, SentiWordNet and SlangSD, performed significantly worse than the others.

5.3 Information Gain

Since the size of the lexicons varied widely, we decided to compare the average words matched per student submission and average sentiment discovered per review. This highlights the information captured by each lexicon. Table 4 presents the average 1) unique lexicon words matched for all 25-35 reviews (Pos/NegWords) and 2) sentiment of key words *per review* (Pos/NegSenti) for both CG and CM.

	CG			CM		
Lexicon	MAE	MdAE	AM	MAE	MdAE	AM
HeLPS_W	**0.115**	0.167	**0.771**	**0.083**	**0.132**	**0.750**
ANEW_W	0.165	0.159	0.514	0.150	0.147	0.417
VADER_W	0.17	**0.157**	0.543	0.150	0.149	0.417
AFINN_W	0.228	0.21	0.571	0.237	0.231	0.472
SWN_W	0.385	0.37	0.2	0.431	0.431	0.111
SlangSD_W	0.574	0.55	0.171	0.591	0.580	0.083

Table 3: Weighted Lexicon Comparison

	CG				CM			
Lexicon	PosWords	NegWords	PosSenti	NegSenti	PosWords	NegWords	PosSenti	NegSenti
HeLPS_W	50.7	7.9	2.345	-0.461	50.4	8.5	2.751	-0.550
ANEW_W	62.7	2.6	2.104	-0.153	69.7	3.9	2.232	-0.229
VADER_W	30.3	4.9	0.775	-0.061	39.4	9.9	0.980	-0.142
AFINN_W	32.8	7.7	0.937	-0.077	38.2	12.9	1.188	-0.203
SWN_W	281.1	76.8	2.437	-1.583	362.8	105.6	2.726	-1.836
SlangSD_W	220.1	145.1	0.325	-0.651	314.4	261.2	0.525	-0.731
HeLPS	*	*	3.332	-0.817	*	*	3.776	-0.903
ANEW	*	*	4.421	-0.327	*	*	4.435	-0.432
AFINN	*	*	2.474	-0.483	*	*	2.706	-0.675
SWN	*	*	6.915	-5.112	*	*	7.635	-5.863
SlangSD	*	*	0.954	-4.289	*	*	1.350	-4.497
MPQA	40.1	19.1	0.251	-0.043	60.2	38.3	0.430	-0.108

* same as weighted version

Table 4: Lexicon Information

HeLPS, ANEW, and SentiWordNet collected the top positive and negative sentiment for weighted lexicons. HeLPS compared favorably with SentiWordNet even though our lexicon contained just 2% and 3% of SentiWordNet's negative and positive words, respectively. ANEW generally found roughly the same positive sentiment, but less negative sentiment than our lexicon (47% and 24% smaller, respectively). As expected, the unweighted lexicons found greater sentiment than their weighted counterparts, since weighting reduced polarity. Of all the lexicons, only SlangSD appeared to be better at finding negative sentiment than positive.

5.4 Qualitative Example

While it is true that a lexicon must find enough sentiment to establish confidence in a grade, the accuracy of the sentiment must also be maintained. Simply matching the most key words is not enough. Figure 3 provides one qualitative example of the three top sentiment-producing weighted lexicons on a single review from a CG group presentation on "Polygon rendering and visible surfaces". The review highlights differences in how the lexicons interpreted text. Blue text denotes words with positive sentiment while red text represents negative

or negated positive sentiment. Black text are words with neutral sentiment.

SentiWordNet matched the most words, but not in an intuitive way (e.g., *mathematical* as positive or *such* and *have* as negative). ANEW's matching was more intuitive, perhaps because its words were selected and weighted by human intelligence, like ours. However, it missed a number of sentiment-bearing words (e.g., *depth* and *clarity*). Ultimately, while all the lexicons found and classified globally positive or negative words (e.g., *attractive* and *complete*), our domain-dependent lexicon correctly captured the words relevant to and important in our context while excluding the noise.

HeLPS_W
Grade: 3.18/4.00 (B)
Topics such as Refraction and Ray tracing illumination were well-explained in depth especially the mathematical concepts and derivations. Abundant inclusion of references and illustrations. I felt that the layout of the presentation slides could have been improved to make it more attractive. Inclusion of an algorithm in the Ray tracing illumination topic provides a complete coverage of topic but could have improved the visible clarity of the algorithm slide.

ANEW_W
Grade: 3.33/4.00 (B+)
Topics such as Refraction and Ray tracing illumination were well-explained in depth especially the mathematical concepts and derivations. Abundant inclusion of references and illustrations. I felt that the layout of the presentation slides could have been improved to make it more attractive. Inclusion of an algorithm in the Ray tracing illumination topic provides a complete coverage of topic but could have improved the visible clarity of the algorithm slide.

SentiWordNet_W
Grade: 2.99/4.00 (B)
Topics such as Refraction and Ray tracing illumination were well-explained in depth especially the mathematical concepts and derivations. Abundant inclusion of references and illustrations. I felt that the layout of the presentation slides could have been improved to make it more attractive. Inclusion of an algorithm in the Ray tracing illumination topic provides a complete coverage of topic but could have improved the visible clarity of the algorithm slide.

Figure 3: Qualitative Comparison of Lexicons

6 CONCLUSION

We outline a data-driven process to score subjective content in CAD course projects using sentiment from peer review comments and evaluate the resulting domain-dependent lexicon against others publicly available. First, weighted lexicons perform better in general quality metrics than do unweighted lexicons. Adding weighting information improves review agreement (lower standard deviation) and score matching between the analytical and subjective portions of the review form (lower mean/median absolute error) in our courses. Secondly, HeLPS, our domain-dependent lexicon, performed concisely and accurately in the CAD course context especially when compared to other publicly available automatic- or hand-ranked lexicons. HeLPS resulted in the lowest difference between aggregate review form score and aggregate comment score and consistently tagged high-quality positive and negative sentiment with a lexicon a fraction of the size of others. Finally, simply matching the most key words or finding the greatest polarity in the text did not guarantee success. A qualitative example demonstrated that a smaller lexicon can outperform a larger one by ignoring noise and increasing domain precision.

In the future, we would like to include another category in our lexicon — intensifiers — which would add more versatility to polarity modification than simple negation. Although our lexicon is currently comprised of single key words (unigrams) we expect increased accuracy by expanding it to include short phrases (bigrams and trigrams). Finally, we are interested in work combining approaches to allow a key word, phrase, or sentence

to be voted on by an ensemble of lexicons. This would allow lexicons with particular areas of expertise (e.g., SlangSD on negative sentiment or emoticons) to balance others and may increase accuracy.

ORCID

Zachariah J. Beasley http://orcid.org/0000-0002-0146-2739
Les A. Piegl http://orcid.org/0000-0003-0629-8496

REFERENCES

[1] Baccianella, S.; Esuli, A.; Sebastiani, F.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. LREC, 10(2010), 2200–2204, 2010.

[2] Balahadia, F.F.; Fernando, M.C.G.; Juanatas, I.C.: Teacher's performance evaluation tool using opinion mining with sentiment analysis. In 2016 IEEE Region 10 Symposium (TENSYMP), 95–98. IEEE, 2016.

[3] Beasley, Z.; Friedman, A.; Piegl, L.; Rosen, P.: Leveraging peer feedback to improve visualization education. arXiv preprint arXiv:2001.07549, 2020.

[4] Beasley, Z.J.; Piegl, L.A.; Rosen, P.: Ten challenges in cad cyber education. Computer-Aided Design and Applications, 15(3), 432–442, 2018.

[5] Beasley, Z.J.; Piegl, L.A.; Rosen, P.: Designing intelligent review forms for peer assessment: A data-driven approach. In 2019 ASEE Annual Conference & Exposition. American Society for Engineering Education, 2019. https://doi.org/10.1080/16864360.2017.1397893.

[6] Belinkov, Y.; Glass, J.: Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7, 49–72, 2019.

[7] Blair-Goldensohn, S.; Hannan, K.; McDonald, R.; Neylon, T.; Reis, G.A.; Reynar, J.: Building a sentiment summarizer for local service reviews. In WWW workshop on NLP in the information explosion era, vol. 14, 339–348, 2008.

[8] Bradley, M.M.; Lang, P.J.: Affective norms for english words (anew): Instruction manual and affective ratings. Tech. rep., Citeseer, 1999.

[9] Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intelligent systems, 28(2), 15–21, 2013.

[10] Chaplot, D.S.; Rhim, E.; Kim, J.: Predicting student attrition in moocs using sentiment analysis and neural networks. In AIED Workshops, vol. 53, 54–57, 2015.

[11] Clarizia, F.; Colace, F.; De Santo, M.; Lombardi, M.; Pascale, F.; Pietrosanto, A.: E-learning and sentiment analysis: a case study. In Proceedings of the 6th International Conference on Information and Education Technology, 111–118. ACM, 2018.

[12] Dave, K.; Lawrence, S.; Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web, 519–528. ACM, 2003.

[13] Deng, L.; Wiebe, J.: Mpqa 3.0: An entity/event-level sentiment corpus. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1323–1328, 2015. https://doi.org/10.3115/v1/N15-1146.

[14] Esparza, G.G.; Díaz, A.P.; Canul-Reich, J.; De-Luna, C.A.; Ponce, J.: Proposal of a sentiment analysis model in tweets for improvement of the teaching-learning process in the classroom using a corpus of subjectivity. International Journal of Combinatorial Optimization Problems and Informatics, 7(2), 22–34, 2016.

[15] Fang, X.; Zhan, J.: Sentiment analysis using product review data. Journal of Big Data, 2(1), 5, 2015.

[16] Fleming, M.; Coulter, B.; Weaver, S.; et al.: Exploring the student experience of industry placements using sentiment analysis. In 29th Australasian Association for Engineering Education Conference 2018 (AAEE 2018), 213. Engineers Australia, 2018.

[17] Haselmayer, M.; Jenny, M.: Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Quality & quantity, 51(6), 2623–2646, 2017. https://doi.org/10.1007/s11135-016-0412-4.

[18] Hatzivassiloglou, V.; McKeown, K.R.: Predicting the semantic orientation of adjectives. In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics, 174–181. Association for Computational Linguistics, 1997.

[19] Hu, M.; Liu, B.: Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168–177. ACM, 2004.

[20] 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.

[21] Lalata, J.a.P.; Gerardo, B.; Medina, R.: A sentiment analysis model for faculty comment evaluation using ensemble machine learning algorithms. In Proceedings of the 2019 International Conference on Big Data Engineering, 68–73. ACM, 2019.

[22] Li, H.; Xiong, Y.; Hunter, C.V.; Guo, X.; Tywoniw, R.: Does peer assessment promote student learning? a meta-analysis. Assessment & Evaluation in Higher Education, 1–19, 2019. https://doi.org/10.1080/02602938.2019.1620679.

[23] Lin, B.; Zampetti, F.; Bavota, G.; Di Penta, M.; Lanza, M.; Oliveto, R.: Sentiment analysis for software engineering: How far can we go? In 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), 94–104. IEEE, 2018.

[24] Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1–167, 2012.

[25] Mudinas, A.; Zhang, D.; Levene, M.: Bootstrap domain-specific sentiment classifiers from unlabeled corpora. Transactions of the Association for Computational Linguistics, 6, 269–285, 2018.

[26] Muhammad, A.; Wiratunga, N.; Lothian, R.: Contextual sentiment analysis for social media genres. Knowledge-Based Systems, 108, 92–101, 2016.

[27] Neviarouskaya, A.; Prendinger, H.; Ishizuka, M.: Sentiful: A lexicon for sentiment analysis. IEEE Transactions on Affective Computing, 2(1), 22–36, 2011.

[28] Ng, A.: Learning from moocs. Inside Higher Ed, 24(1), 2013.

[29] Nielsen, F.Å.: A new anew: Evaluation of a word list for sentiment analysis in microblogs, 2011. https://arxiv.org/abs/1103.2903.

[30] Novielli, N.; Girardi, D.; Lanubile, F.: A benchmark study on sentiment analysis for software engineering research. In 2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR), 364–375. IEEE, 2018.

[31] Pang, B.; Lee, L.; Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 79–86. Association for Computational Linguistics, 2002.

[32] Pang, B.; Lee, L.; et al.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135, 2008.

[33] Piegl, L.A.; Beasley, Z.J.; Rosen, P.: Assessing student design work using the intelligence of the crowd. In Proceedings of CAD'19, 117–121. CAD Conference and Exhibition, 2019. https://doi.org/10.14733/cadconfP.2019.117-121.

[34] Pousada, M.; Caballé, S.; Conesa, J.; Bertrán, A.; Gómez-Zúñiga, B.; Hernández, E.; Armayones, M.; Moré, J.: Towards a web-based teaching tool to measure and represent the emotional climate of virtual classrooms. In International Conference on Emerging Internetworking, Data & Web Technologies, 314–327. Springer, 2017.

[35] Rani, S.; Kumar, P.: A sentiment analysis system to improve teaching and learning. Computer, 50(5), 36–43, 2017.

[36] Ribeiro, F.N.; Araújo, M.; Gonçalves, P.; Gonçalves, M.A.; Benevenuto, F.: Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1), 23, 2016.

[37] Schubert, M.; Durruty, D.; Joyner, D.A.: Measuring learner tone and sentiment at scale via text analysis of forum posts. In Proceedings of the 8th Workshop on Personalization Approaches in Learning Environments (PALE 2018). Kravcik, M., Santos, OC, Boticario, JG, Bielikova, M., Horvath, T. and Torre I.(Eds.). 19th International Conference on Artificial Intelligence in Education (AIED 2018), CEUR workshop proceedings, this volume, 2018.

[38] Serrano-Guerrero, J.; Olivas, J.A.; Romero, F.P.; Herrera-Viedma, E.: Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18–38, 2015.

[39] Snow, R.; O'Connor, B.; Jurafsky, D.; Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the conference on empirical methods in natural language processing, 254–263. Association for Computational Linguistics, 2008. https://doi.org/10.3115/1613715.1613751.

[40] Soroka, S.; Young, L.; Balmas, M.: Bad news or mad news? sentiment scoring of negativity, fear, and anger in news content. The ANNALS of the American Academy of Political and Social Science, 659(1), 108–121, 2015.

[41] Taboada, M.; Brooke, J.; Tofiloski, M.; D. Voll, K.; Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics, 37, 267–307, 2011. http://doi.org/10.1162/COLI_a_00049.

[42] Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267–307, 2011.

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

[44] Thelwall, M.; Buckley, K.; Paltoglou, G.; Cai, D.; Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558, 2010.

[45] Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, 417–424. Association for Computational Linguistics, 2002.

[46] Wen, M.; Yang, D.; Rose, C.: Sentiment analysis in mooc discussion forums: What does it tell us? In Educational data mining 2014. Citeseer, 2014.

[47] Wiegand, M.; Balahur, A.; Roth, B.; Klakow, D.; Montoyo, A.: A survey on the role of negation in sentiment analysis. In Proceedings of the workshop on negation and speculation in natural language processing, 60–68, 2010.

[48] Wilson, T.; Wiebe, J.; Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005.

[49] Wu, L.; Morstatter, F.; Liu, H.: Slangsd: Building and using a sentiment dictionary of slang words for short-text sentiment classification, 2016. https://arxiv.org/abs/1608.05129.