Title:

HeLPS: A Domain-Specific Lexicon for CAD Peer Review

Authors:
Zachariah J. Beasley, zjb@mail.usf.edu, University of South Florida
Les A. Piegl, lpiegl@gmail.com, University of South Florida

Keywords:
Sentiment Analysis, CAD, Lexicon, Peer Review, Educational Data Mining

DOI: 10.14733/cadconfP.2020.21-25

Introduction:

Grading creative works requires a subjective component. 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 [8], [9]. 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 [2], [11].

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. 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. [6]) that the process, rather than the lexicon itself, should be copied.

The main contribution of this paper is the introduction of a process to create a domain-dependent lexicon from student peer review text, implemented specifically in a CAD-course context and compared to other publicly available lexicons. The structure of the courses ([11]), design of the review form ([3], [11]), and implementation of the sentiment algorithm ([3]) are covered more fully in prior work and thus only briefly mentioned when relevant.

Process:

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) [11]. Over the course of six semesters (eleven courses, both CAD and non-CAD with approximately 425 students), we gathered any sentiment-bearing key words from the reviews — even those used very infrequently — to add to

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

Tab. 1: Sample of Lexicon Words

our lexicon, HeLPS: the Heuristic Lexicon of Peer Sentiment. 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. Table 1 shows some sample key words from our lexicon.

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 (left) 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 (left) exemplifies both general (e.g. *understand* and *useful*) as well as domain-specific key words (e.g. *cite* and *diagram*).

Polarity:

The polarity of a word, phrase, or sentence is comprised of direction (positive, negative, neutral) and an optional weight. 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 1 (right) demonstrates how each set fit into three sentiment directions — positive, negative, neutral. There was overlap between negative sentiment and negate words (e.g. *missing*) as well as between negative sentiment and flag words (e.g. *copying*). Reset tokens were always neutral (e.g. *however*).

After direction of sentiment was determined, key words were weighted by instructor heuristic as opposed to a learning strategy like [1] and [14]. Other lexicons are similarly weighted by an expert [10], small group [4], or a crowd (e.g. Amazon Mechanical Turk workers in [7], [6], and [12]). For simplicity, both positive and negative key words were weighted on a continuous scale between zero and one.

Experiment:

We compared the aggregation of sentiment between our lexicon and six others publicly available — AFINN-111 [10], ANEW-2017 [4], MPQA [5], SentiWordNet 3.0 [1], SlangSD [14], and Vader [7] — 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 [3] and [11] for details and [13] for a similar negation strategy). The 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 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 comment score to their corresponding review form analytical score, rather, the aggregation of each from 25-35 reviews. We contrast the following attributes on our two most recently completed mixed graduate and undergraduate CAD courses, Computer Graphics (CG) and CAD Modeling (CM):

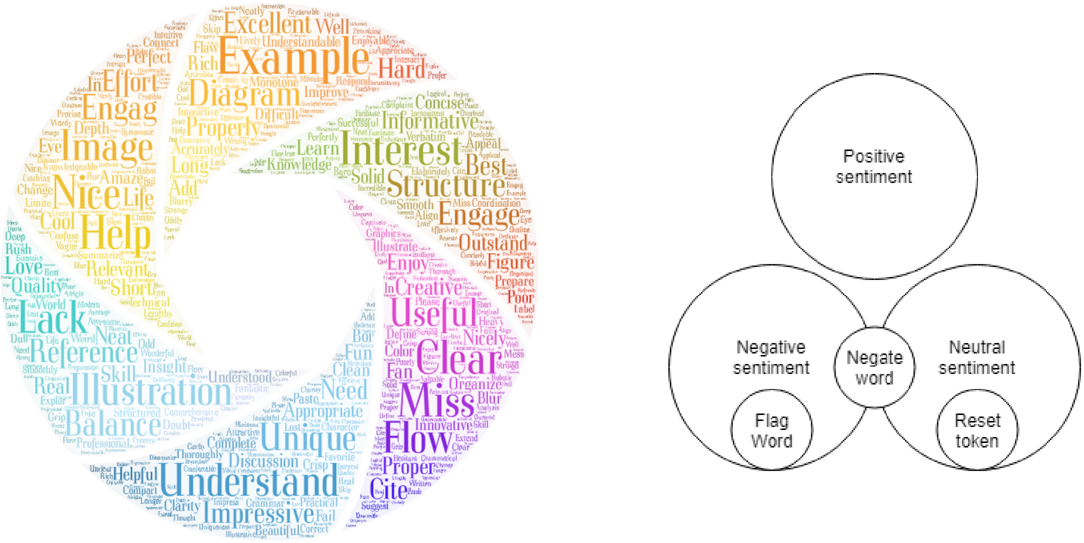


Fig. 1: Left: Word Cloud of Semester Key Words; Right: Polarity of Tokens

	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

Tab. 2: Weighted Lexicon Comparison

- 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)

Table 2 shows the lexicon mean/median absolute error and average matched, with the best scores bolded for the weighted lexicons. 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. The two largest lexicons, SentiWordNet and SlangSD, performed significantly worse than the others.

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 3 presents the average 1) unique lexicon words matched for all 25-35 reviews

	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

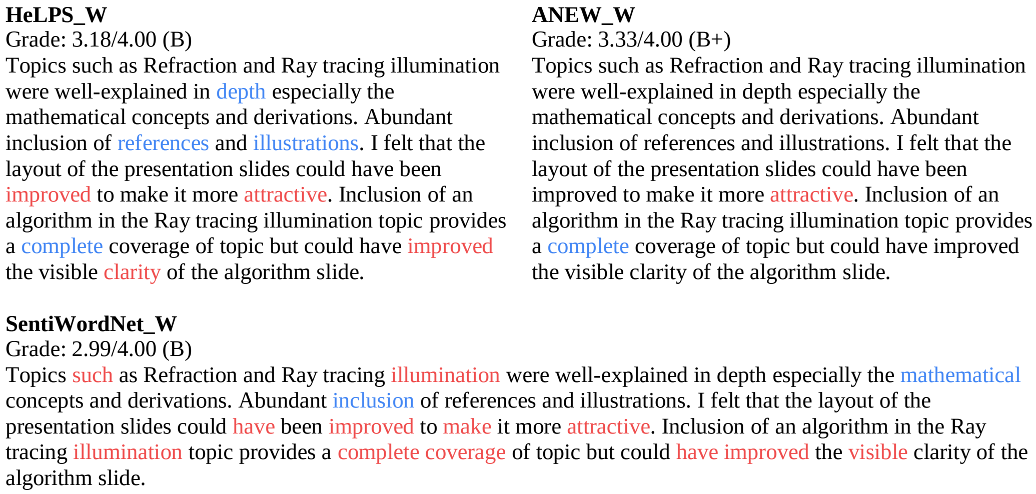
Tab. 3: Lexicon Information

(Pos/NegWords) and 2) sentiment of key words *per review* for both CG and CM.

HeLPS, ANEW, and SentiWordNet collected the top positive and negative sentiment. 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). Of all the lexicons, only SlangSD appeared to be better at finding negative sentiment than positive.

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. Figure 2 provides one qualitative example of the three top sentiment-producing 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 negated positive or negative 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 most relevant and important words 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.

Fig. 2: Qualitative Comparison of Lexicons

Conclusion:

We outline a data-driven process to score subjective content in academia using sentiment in peer review comments. 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 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.

References:

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

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

[3] 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.

[4] Bradley, M. M.; Lang, P. J.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[5] 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. 2015, 1323–1328. https://doi.org/10.3115/v1/N15-1146.

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

[7] 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.

[8] 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, 2019, 1–19. https://doi.org/10.1080/02602938.2019.1620679.

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

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

[11] Piegl, L. A.; Beasley, Z. J.; Rosen, P.: Assessing Student Design Work using the Intelligence of the Crowd. In Proceedings of CAD'19. CAD Conference and Exhibition, 2019, 117–121. https://doi.org/10.14733/cadconfP.2019.117-121.

[12] 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. Association for Computational Linguistics, 2008, 254–263. https://doi.org/10.3115/1613715.1613751.

[13] 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. 2010, 60–68.

[14] 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.