Introduction

At some point, many evaluators are faced with a challenge of "getting it done well" versus "getting it done on time." Evaluators often compromise some part of a study in order to be thorough in another due to time restrictions. As a result, practitioners can be confronted with the less than ideal choice between reporting on a partially analyzed data set or extending a study beyond its intended length. This may be most apparent when time constraints are put on the analysis of open-ended responses derived from interviews, surveys, etc.

Qualitative research methods provide the seminal tools for analyzing open-ended text data. However, the amount of resources needed to process these data is often dependent on the availability of time, amount of training and/or experience, and the number of qualitative analysts. Depending on the size of a dataset, the sheer quantity of open-ended text may force researchers to prioritize some aspect of a study over others, leaving some amount of potentially groundbreaking data in its raw form. In instances when timely feedback is important, the data are plentiful, and answers to the study questions carry lower consequences, Artificial Intelligence (AI) is a promising and nascent way to efficiently mine data to answer evaluation questions.

Though researchers traditionally use qualitative methods as a means of discovering underlying concepts and themes hidden within narratives, we make a case for the use of a quantitative technique, namely the use of text mining (TM) and machine learning (ML) to analyze text data as a means to discovering underlying sentiments and themes. In this manuscript, we begin with an accessible primer on TM and ML including relevant limitations, vocabulary, and an introduction to a specific yet highly applicable ML approach known as

sentiment analysis. Finally, we provide an example evaluation case to highlight the utility of such an approach. Before proceeding, the authors would like to stress that text mining techniques are not a replacement of qualitative methodology, which remains the gold standard for analyzing open text.

Natural Language Processing: A Primer

While many technical explanations of ML exist, this field can be described in simple terms as a subset of AI with a focus on the *science of building an environment where a computer can perform some actions without being explicitly programmed to do so* (Samuel, 1959). In practice, ML presents a varying and wide-ranging choice of techniques to categorize text data and to make predictions on those classifications. This is done by partitioning an existing data set into subsets that are for testing and training. When assessing classification, predictive accuracy is typically estimated using various statistical techniques including sensitivity, specificity, and accuracy. This diverse set of tools makes the field highly intriguing, even by those who have limited to no methodological knowledge of ML. Unfortunately, the field of ML and the limitations of AI-driven analysis can be misunderstood and distorted by individuals with limited understanding.

The Limits of Artificial Intelligence

Humanlike characteristics. Human characteristics are often wrongly ascribed to AI. In practice, mechanical "brains" can perform tasks such as making decisions, learning content, memorizing materials, and forming predictions within the scope of their underlying code. While this list may appear similar to a set of activities humans do, the primary difference lies in the fact that machines cannot perform any task beyond the scope of their underlying programming.

In contrast, humans generally tend to grow beyond the scope of their "programming" throughout their lives.

Differences in Languages. The primary difference between human and computer languages lies in the structural disparities found in the morphology of each. Although the human language and its set of rules have developed and are continuing to evolve *naturally* over time. Computer languages, however, are built to be contained, static, and well-defined. Without assistance from people, a machine needs to find other means of support when analyzing, comprehending, and deriving meaning from text using an intelligent and convenient approach. A popular technique is to apply natural language processing (NLP), which uses statistical algorithms to recognize, classify, and extract particular linguistic rules so a machine can understand written text created by humans. Broadly speaking, these rules attest to specific grammatical structures known as *syntax* and *semantics* (Bates, 1995).

Syntax refers to the organization of terms in a sentence so that it is grammatically sound whereas *semantics* denotes the meaning underlying a body of text. In *syntactic analysis*, a machine uses the arrangement and sequencing of words to determine their alignment with grammatical rules. An NLP then uses *semantic analysis* by applying probability-based statistical algorithms to groups of terms, resulting in an ability to interpret and assess meaning. Semantic analysis is difficult because it requires computers to learn with few instructions. Additionally, the use of ambiguous language, common sense knowledge, and the use of symbols complicate analyses as well. For this reason, results from our semantic analysis are beyond the scope of this paper.

Sentiment Analysis

Identifying the emotional state of a set of participants is one of the many benefits of using qualitative methods to analyze open-ended text data. Historically, thematic analysis has been the most widely used technique to classify, study, and report patterns found within a text. If applied correctly, a rigorous thematic analysis with proper checks and balances can produce trustworthy and revealing findings (Braun & Clarke, 2006). However, a difficulty associated with this method is that the approach must be applied in a very specific manner, and any deviation may result in inconsistencies in the development process of themes (Holloway & Todres, 2003). Moreover, a second disadvantage, as denoted by Braun and Clarke (2006), is that thematic analyses do not provide a foundation for researchers to make claims regarding language use (Könings et al., 2011). While not perfect, the complementary quantitative approach to sentiment analysis is an option for analyzing text for emotion.

Sentiment analysis, also known as opinion mining, is the automated task of determining what feelings a participant is expressing in text; typically framed as the binary distinction of positive and negative attributes (Zhang & Liu, 2017). However, this type of analysis is also used in a more granular approach when the primary objective is assessing the emotions of an individual. This is accomplished by assigning terms discrete scores, binary measures of negative/positive polarities, or simple emotional states (e.g. anger, fear, or sadness).

There are two major approaches to conducting this mode of analysis. The first requires *supervised learning,* where pre-labeled data are used to train a machine. The second is known as *unsupervised learning* and relies on an external reference, such as a lexicon, to perform analysis. *Lexicons* are special dictionaries that contain a list of terms and their negative or positive polarities provided by a scoring system. Predicated on the lexicon used, this score is

dependent on any number of conditions such as a term's (a) context, (b) direct connectivity to other words (such as those that precede and follow it), and (c) position.

There are multiple ways to perform a lexicon-based sentiment analysis but most, if not all approaches, use the following general steps: (a) construct or use a predefined lexicon; (b) aggregate the number of positive and negative sentiments; and (c) assess groups of negative and positive words to find clusters that are either mostly negative or positive. Next, we walk through an example of how the steps explained above are applied to a set of responses to three assignments used in an evaluation.

General Approach

While a specific set of tasks leading to the sentiment analysis are outlined below, in the next section we provide a detailed visualization of the process taken in Figure 2. Without using an abundance of terminology specific to ML, we outline the general procedure used in conducting our sentiment analysis:

Table 1

Process in Conducting a Sentiment Analysis

ML Stage	Description	General Steps Taken
Preprocessing and *feature engineering*	Preliminary step where raw text is put into a form ready for analysis.	1. Text data are cleaned and put into a form that a machine can understand and use. 2. Words are put into single column vectors.
Training parameters and *testing*	Analysis of text is performed by splitting the existing set or words into a training group and a testing group such that the latter is	1. The text data are split into two groups: a training set with 20% and a test set with

	a. large enough to produce statistically significant results, and b. terms are generally representative of those used throughout the entire text data set.	80% of the terms, respectively[1]. 2. Sentiments are assessed using NLP and lexicons. 3. A predictive model is generated that is then applied to the test set.
Context assessment and reclassification	Post analysis step where terms whose sentiment classification may be incorrect or uncertain are visually inspected and if needed, rescaled or reclassified.	1. Terms and associated sentiments that have a lower likelihood of being true as determined by outliers within the model are observed in context. 2. Changes to the sentiment or its value are made if necessary.

The three stages are by no means unique, in that NLP provides a means for translating words and phrases and is essentially a broad framework consisting of numerous techniques to translate and handle human language. Specific tasks and what order they should be administered when applying an NLP do not exist for the primary reason that there are multiple approaches that a person can take when mining and analyzing written text. Next, we describe a set of standard text cleaning steps used in the first stage.

Tokenization and Lemmatization. A highly utilized approach when preparing terms is in the use of *tokenization*, a process by which text is separated into pieces consisting of characters, known words, phrases, segments, etc. This is followed by the deletion of *stop words*, or commonly used terms (e.g. “a”, “an”, “the”, etc.; Grefenstette & Tapanainen, 1994).

[1] Exemplifying the Pareto Principle, the 80-20 split is standard practice from the view that an accurate model fit is influenced by more training data resulting in a reduced variance in the results, in that approximately 80% of a model’s effect comes from about 20% of the causes (Kilicoglu et al., 2019). However, readers should be aware that this is simply an estimate and the split is contingent on a given data set.

Next, *lemmatization* is applied by categorizing the inflected forms of a term so they can be analyzed as a single word (e.g. “engineer” and “engineers” get grouped into “engineer”; Karlsson, 1994). Finally, a representation of text that indicates the occurrence of remaining terms within the document, or a *bag-of-words* (BoW; Goldberg, 2017, pp. 67, 187-189), is constructed. The result in this preprocessing state is a two-column matrix with terms and corresponding frequencies, respectively.

Text Classification. Assigning text to one or more categories of natural language documents is known as *text classification*. The Naïve Bayes classifier is a probabilistic learning model predicated on Bayes’ Theorem that assigns corpora and assumes all features (e.g. context, lemmatization, tokenization, etc.) are independent of one another between categories (Russell et al., 2010, pp. 495-499). In everyday usage, this process is applied in the filtering of spam email or routing of customer support calls. However, in cases related to text mining, the Naïve Bayes’ Classifier is used to assess a writer’s point of view or to predict key topics about activities, products, services, etc. In particular, the probabilistic classifier serves as a model to train a machine to group bodies of text. To illustrate this model, we provide an example in the next section.

Contextual Example

Introduction

Engineering and computer science are historically male dominated fields. In 2015, women earned just over 20% of all the bachelor’s degrees in engineering (Roy, 2018). In 2017, women earned approximately 19% of all bachelor’s degrees in computer science (Trapani & Hale, 2019). Historically only 30% of women who enter engineering are still working in

engineering 20 years later (Corbett & Hill, 2015). Additionally, even though men and women have been shown to have equivalent grade point averages, women consistently report lower levels of self-efficacy in engineering and higher levels of discrimination than men (Christina et al., 2007). In particular, undergraduate engineering classes and the engineering profession have a reputation for being a chilly environment for persons who identify as women (Allan & Madden, 2006; Hall & Sandler, 1982). And in a qualitative study of women in science and engineering (*n*=26), Hughes (2012) noted the women who were most likely to leave science and engineering were those who either endorsed gender stereotypes or saw themselves as more feminine than typical members of those fields. While substantial efforts are underway to support women in undergraduate engineering and computer science classrooms, there is still much to be done to change the undergraduate classroom climate to be more welcoming of women.

As part of a National Science Foundation (NSF) grant (omitted for blind review), the research team for the program being evaluated developed and implemented multiple activities into several engineering and computer science classes at three universities to address this chilly climate. Specifically, the purpose of the activities was to help students develop inclusive professional identities in an attempt to change the climate of undergraduate engineering programs (Authors, 2018). The team defines inclusive professional identities as students who recognize and seek out diversity in their teams, work in teams to capitalize on the diversity present to strengthen their teams and relevant outcomes and consider a broad range of potential consumers when designing products or services (Authors, 2017). Part of a larger process evaluation of the program, this study examines the potential differential impact of

three assignments developed for use in a second-semester first-year engineering course at one of the campuses. Students responded to a common set of open-ended reflection items for each assignment. The evaluation team was charged with, among other things, providing some feedback on the activities the research team created for the project. Here we describe the ways in which students who identify as women and those who identify as men responded to three such assignments and our recommended changes to the assignments.

Methodology

***Population and Sample*.** While three universities participated in the NSF grant-funded activities, we examine data selected from only one of the participating institutions, namely a large land grant institution in the eastern US. Students at this university complete a common set of first-year courses before moving into their specific engineering major. The common first-year includes foundational courses in engineering spanning two semesters where activities related to culture, diversity, and/or teamwork in the context of computer programming were implemented.

At this university alone, more than 20 sections of engineering classes participated in some capacity in the grant activities or served as baseline sections each semester. The study was approved by the university IRB and informed consent was collected via an electronic survey at the beginning of the semester. The number of students participating in any given semester is well over 1,000 with an average of three activities each, for approximately 72,000 individual responses to questions just at this campus (1,000 students × 3 activities × 4 questions × 6 semesters of intervention activities). In the absence of a large team of qualitatively trained

evaluators, the vast majority of data would go unanalyzed due to the sheer volume of data collected.

Our sample for this study consisted of students enrolled in a second-semester foundational engineering course in spring 2018, which has no prior computer programming prerequisite. The students responded to a set of reflection questions for each of the grant developed activities. Of the 43 students who provided consent to participate in the research study, 93% were first-year students, 98% indicated they were White, and 61% self-identified as male and 39% self-identified as female[2]. From those, 40 students (24 males, 16 females), 36 (21 males, 15 females), and 42 (25 males, 17 females) completed the three intervention activities and reflection questions. The data consisted of 158 different open-ended response sets. Notably, the data presented here represents a small slice of the larger study and related data collected.

For a general understanding of the cost versus benefit associated with running a sentiment analysis in comparison to a conducting qualitative study, approximately ten hours were dedicated to writing the machine learning program in R with less than one minute needed in total to run the analysis and visualize corresponding results for all response sets. In comparison, Miles et al. (2014, p. 52) estimated that roughly between two to four days were needed to qualitatively code and report on a single case of open ended text. Assuming an eight-hour workday and considering that a case may reasonably be an entire response set associated

2. Of note, the question asked students to please identify their sex and students were given the choices: male, female, I do not identify as either. I identify as: (insert choice), and Prefer not to respond. All students either selected male or female. Of note, we recognize that sex and gender are distinct constructs and not synonymous with each other. Because of how the students were asked to respond, we refer to students as male or female and not men and women.

with one of the assignments, the labor necessary to produce credible and dependable results is estimated to be between 6 - 12 days for these data. This fits the criteria of instances where timely feedback is important, the data are plentiful, and the results —while informative— are not highly consequential. The questions answered by this example sentiment analysis are related to improving assignments for future classes and may be used repeatedly in its current form without additional time for analysis. Now that the code has been developed, the only potentially time consuming task for future assignments is simply in prepping the data, student responses, for analysis. Thus, the real benefit of NLP is maximized when repeating this analysis on future datasets.

***Description of Assignments*.** Three assignments were incorporated into this course to address issues of culture, diversity, and/or teamwork. The assignments and all common reflection questions are available at http://partnership4equity.org/resources.html. The reflection questions are prime for using sentiment analysis because the same reflections questions were used on each assignment, and the students responded with open-ended text.

For this example, we chose the response sets from one of the common reflection items due to the potentially contentious nature of the question and the anticipated possibility of polarizing viewpoints. This question asked respondents to describe what they learned about working on teams with other engineers and non-engineers and how that information could make them a better team member.

Algorithmic Justice League. This online homework assignment focused on the development of software and interfaces for diverse populations. The students watched a video where the speaker, a computer programmer who was a female African American, explained

how implicit biases appear to be programmed into code. She uses an example detailing how initial facial recognition software was developed by light-skinned people and was unable to identify people with very dark skin.

Neuroplasticity. This online homework assignment focused on neuroplasticity. The students watched a video about how the brain performs like a muscle, and the more it is used and practices, the better it works. The video also explained how people learn at different rates and have different learning preferences. Students were expected to learn the need to: (a) persist in difficulty and (b) acknowledge teammates are likely to approach the tasks with a diversity of readiness based on their prior experiences.

Wage Gap. This MATLAB® homework assignment discussed the 4% wage gap between men and women in engineering careers (averaged across disciplines). Students applied computer programming skills learned in the course such as loops, user input, plotting, and generating tables to create a computer program that analyzed the wage gap. After each assignment, students completed reflections questions in addition to content-based questions.

An Analytic Approach. While there are numerous lexicons available to use for any sentiment analysis, we chose three that were well-known, analyze text using different measures, and have been tested for validity and reliability (e.g. Khoo & Johnkhan, 2018; Reagan et al., 2017; Weissman et al., 2019). All of the selected lexicons use single words for their base measures and uses associated positive/negative scales: (a) AFINN (Årup Nielsen, 2011) which measures the severity of positive or negative terms using an integer a scale of -5 to 5, (b) bing (Ding et al., 2008) that only categorizes the polarity of a word, and the (c) NRC Emotion (Mohammad & Turney, 2013a, 2013b) which has the most versatile categorization with not

only partitioning wording into positive and negative, but also classifying them into emotional states including anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. These dictionaries are well-cited within the ML literature. The entire list of terms and sentiments are publicly available on the software development platform GitHub, and they use different measures which were used for term and sentiment corroboration. In addition, the three lexicons serve as a way to triangulate the results. Each lexicon relies on different lists of words and ways to characterize the BoW.

Design. The researchers used a process similar to that outlined by Silge and Robinson (2017) and shown in Figure 1.

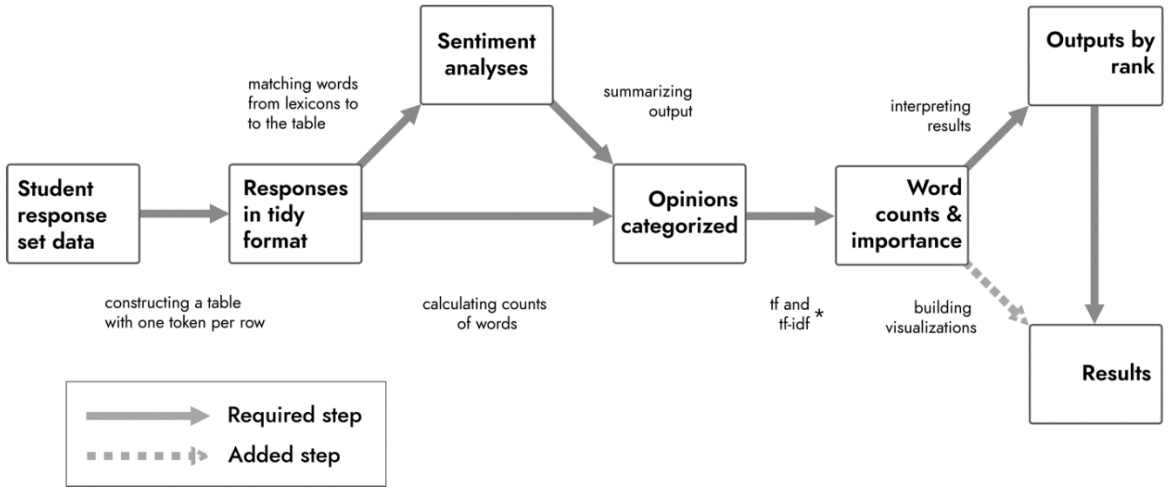

Figure 1. General process taken for conducting a sentiment analysis within a study.

* denotes the implementation of statistical measures of term inverse document frequency (*td-idf*) used to measure a word's importance (Rajaraman & Ullman, 2012, p. 8).

This modified framework was chosen for two reasons. First, this framework is one of many established approaches and, second, can be implemented using the freely available software

package R (R Core Team, 2014), which we used for all analyses. Text mining and lemmatization was conducted using the R packages tm (Feinerer & Hornik, 2018; Feinerer et al., 2008) and textstem (Rinker, 2018), respectively. All data loading, management, structuring and visualization was conducted using the tidyverse family of packages which includes ggplot2, dplyr, tidyr, purrr, tibble, stringr and forcats (Wickham, 2017) and lubridate (Grolemund & Wickham, 2011). In conducting the sentiment analyses, we utilized the package tidytext (Silge & Robinson, 2016) which has a framework consistent with the tidy environment and relies on the packages tokenizer (Mullen et al., 2018) and quanteda (Benoit et al., 2018) for quantitatively analyzing the text data, respectively.

Process. As noted above, there are multiple approaches to derive sentiments. For simplicity, we have categorized the procedure into three stages (Figure 2) with further details provided about the matching process between an open set of terms and those within the lexicons (Figure 3).

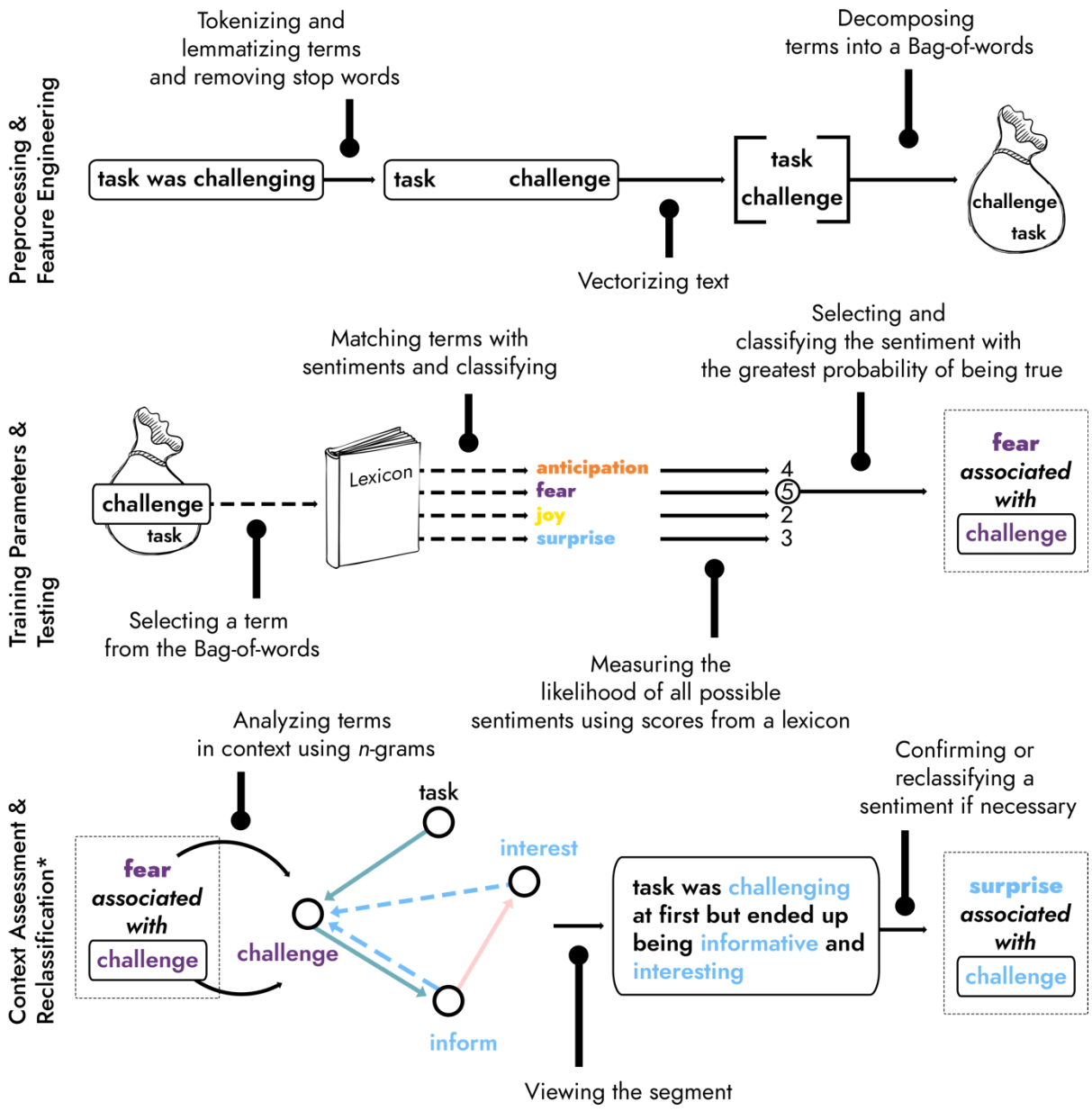

Figure 2. Generalized method for conducting the sentiment analysis used within the study. The graphic uses the NRC lexicon to illustrate the process applying it to a sentence segment from the open-ended responses data set. A detailed view of the selection and matching approach can be found in Figure 3.

In the Preprocessing & Feature Engineering phase, the tokenization and lemmatization of terms occurred, thus reducing each word to its base form without any inflections. Moreover,

the list of terms was filtered against existing stop words reducing the overall number of expressions. These were then grouped into a one-dimension vector and processed into a BoW losing their original order as a result. While we lost information regarding the sequencing of each term, there were significant gains in processing speed (Wang & Manning, 2012).

In the Training Parameters & Testing phase, we retained terms from the BoW that were tagged with a positive or negative sentiment consistently across the three lexicons. Afterward, negation and intensification scores were applied, particularly in assigning AFINN severity scores, Bing identifiers as positive, negative, or neutral, and the NRC classes of anger, anticipation, disgust, fear, joy, sadness, surprise, or trust.

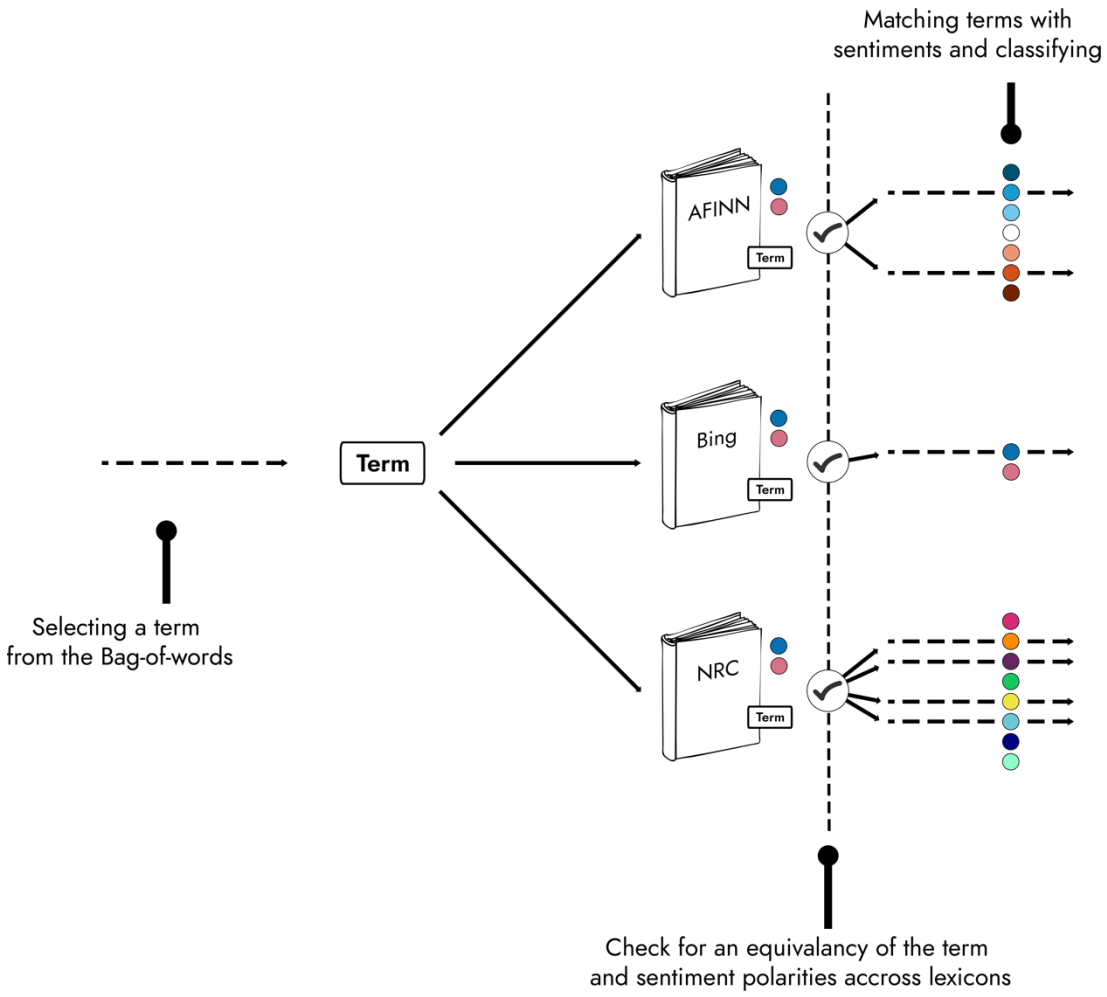

Figure 3. A closer look at inclusion criteria shows that a term from the BoW must be referenced in all three lexicons and have consistent positive and negative sentiment values. To uncover a consistency in polarities, Bing and NRC match terms using these polarities by default. In the case of applying the AFINN Lexicon, we collapsed the negative and positive values to represent these polar categories,

Finally, for those instances where the methodological triangulation fails to accurately describe a sentiment, we used bigrams (a sequence of two consecutive words) and trigrams (a sequence of three consecutive words) to reduce the implied assumption of independent words used in the BoW model. In these circumstances, we retained coordinating, correlative, and subordinating conjunctions (e.g. *and*, *either/or*, and *even if,* respectively) while removing all other stop words.

With the use of network analysis, the entire body of text with conjunctions was mapped whose outcome is a collection of bigrams and trigrams, or a subnetwork of two or three nodes connected by one or two edges, respectively (See Figure 4). For example, in the example network displayed in the last row of Figure 4, the sequence of terms *task*, *challenge*, and *inform* represent a known trigram because this ordered triplet can be observed within the feedback data.

Figure 4. General structure of a digraph (left) and trigraph (right).

Moreover, allowing conjunctions to remain in our data set allowed us to observe indirect triads. Using Figure 2 once again, the pink directed edge indicates that the terms

inform, and *interest* are connected by a conjunction, which in this case was the word *and*. This conjunctive term indicated that the sentiment of *challenge* may need to be updated. Because both *interest* and *inform* were strongly associated with surprise, the sentiment associated with *challenge* was updated whose effect is given by the light blue dotted directed edges. This was verified upon inspection of the entire sentence in the original feedback data set.

Results

Activity-level sentiments were found by assessing the common terms between lexicons and were ranked by term document-inverse document frequency (*td-idf)* measures to signify term importance. To summarize our approach, we first (a) prepared the data by removing all of the stop words, (b) performed tokenization and lemmatization of terms, (c) vectorized the simplified set, (d) reduced the complexity of the data set using *td-idf*, which both produces numerical features for classification and decreased the importance of common terms, (e) created a training set consisting of 80% of the data which was tagged by three common lexicons, (f) recursively amended as needed using *n*-grams, (g) tested that set on the remaining 20% of the data, and (h) derived the final sentiments by using information from all three aforementioned lexicons. Since our intent was to explore the sentiments delineated by sex rather than to conduct a comparative study between each group, when reviewing the visualizations of outcomes, the reader should look at both sets of findings independently. The results are described below.

Algorithmic Justice League. The only term common across males and females was "change" and was indicated by NRC. Both males and females noted "change" multiple times and with a fearful undertone. Across all three lexicons, females noted "bias" with high

importance; but for male students, the term “bias” did not make the top 10% on any lexicon. Also, for female students, the emotions associated with “bias” were consistently negative. In contrast, “mistake” was the only term consistently indicated by all three lexicons for male students, and male students were generally surprised by the type of errors that occurred in the coding of the project.

Based on the results, we made several recommendations. One was to be more specific about the learning targets. For example, in the Algorithmic Justice League assignment, identifying and correcting bias in coding was one of the main objectives, but the male students failed to mention bias in the top 10% of words used when responding to the question asking them what they learned from the activity. The original prompt in the activity asked students to “identify two specific elements in the National Society of Professional Engineers (NSPE) code of ethics that may have been violated in the scenario.” To be sure to draw attention to the underlying issue of bias in the coding, we recommended the activity prompt should be revised to include: “identify at least two elements of the NSPE code of ethics that were violated *by the bias in the underlying code* in the scenario”.

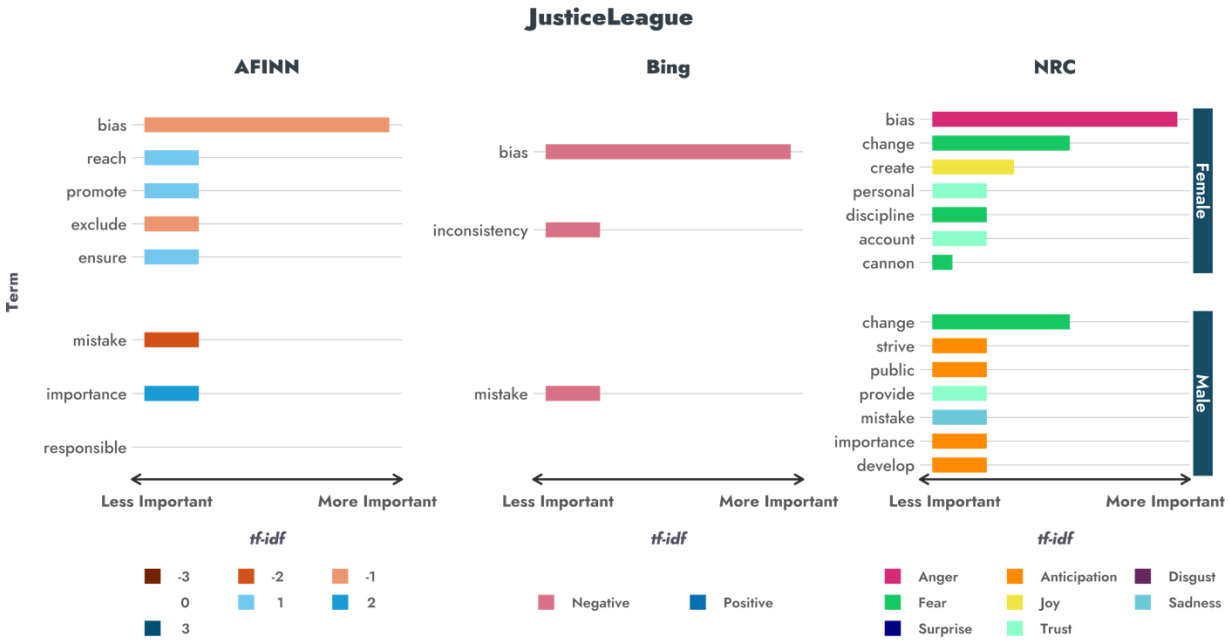

Figure 5. Sample of top 10% of common terms across all Algorithmic Justice League activities ranked by *td-idf*. Terms with empty bars indicate a neutral sentiment.

Neuroplasticity. Both groups of students described fears associated with the task or a result thereof. Males and female students both noted fear associated with "change" but the emotion was stronger for male students than female students. For female students, the terms "easy" (associated with positive emotions) and "difficult" (associated with negative emotions) appeared in multiple lexicons. Female students also noted "difficulty", often and with fear and negative emotions, while male students' opinions suggested ease and success as noted by the terms "successful", "improve", and "safe", all associated with positive emotions and trust. However, male students also indicated "struggle" and "challenge", which were associated with negative emotions and fear.

The neuroplasticity activity was less polarizing for male and female students than the algorithmic justice league assignment. The activity appeared to elicit the types of emotions that would be expected from the assignment. We recommended to the researchers they may want

to consider changing the first question on the activity from “How do you see the relationship between struggle and learning?” to “How is the relationship between struggle and learning potentially both a positive and a negative one?” The revised prompt requires students to move from just basing their answer on how they “see” the relationship between struggle and learning to how the relationship is complex and normalizes struggle in the context of learning.

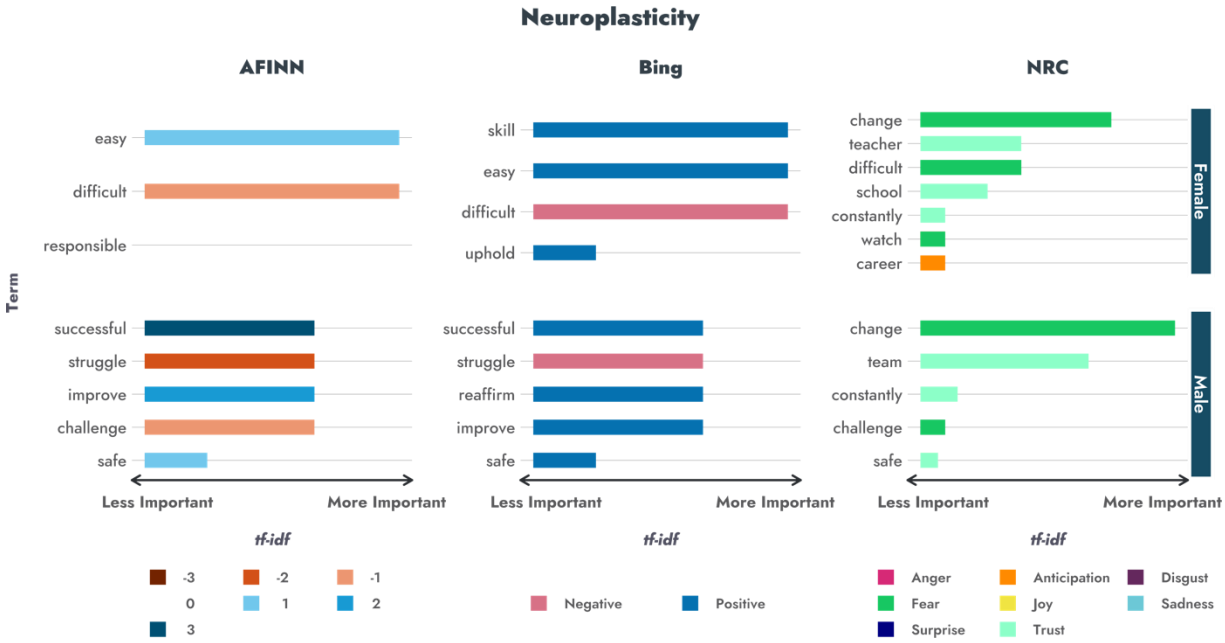

Figure 6. Sample of top 10% of common terms across all Neuropasticity activities ranked by *td-idf*. Terms with empty bars indicate a neutral sentiment.

Wage Gap. Both female and male students associated the term “pay” with both slightly negative emotions (AFINN) but also with positive emotions such as trust (NRC). Additionally, neither female nor male students were angered nor surprised about the result, which may indicate they were generally either already aware of the disparity within engineering or it was simply a facet of today’s society. Also, both male and female students noted words like “discrimination” and “discriminatory” and had strong negative associations. Female students also had a negative response to the findings as illustrated by the high frequency of “change”

and its association with fear and realization. Societal bias (as seen by the terms "discrimination", "inequality", and "change" all associated with negative emotions) was a consistent theme amongst female students while their counterparts appeared to imply that strides were being made towards equality (as seen by "safe" and "fair" associated with positive emotions, and "pay" and "treat" associated with the emotion trust) and may still need to be addressed (as noted by "concern", "issue", "discriminatory" and "blame" associated with negative emotions). Of note, "responsible" was in the top 10% of words for male students but there was no emotion associated with the term.

Finally, like the algorithmic justice league assignment, the wage gap assignment also elicited strong emotions from both male and female students. Our recommendation to the researchers was to add an item to ensure students view they can be part of the solution for the gender wage gap in engineering. For example, the assignment could close with a prompt of "How problematic is the entry-level wage gap on a woman's lifetime earning potential? Based on the examples of companies that have successfully closed the gender wage gap and your own ideas, what can be done to mitigate existing gender wage gaps?"

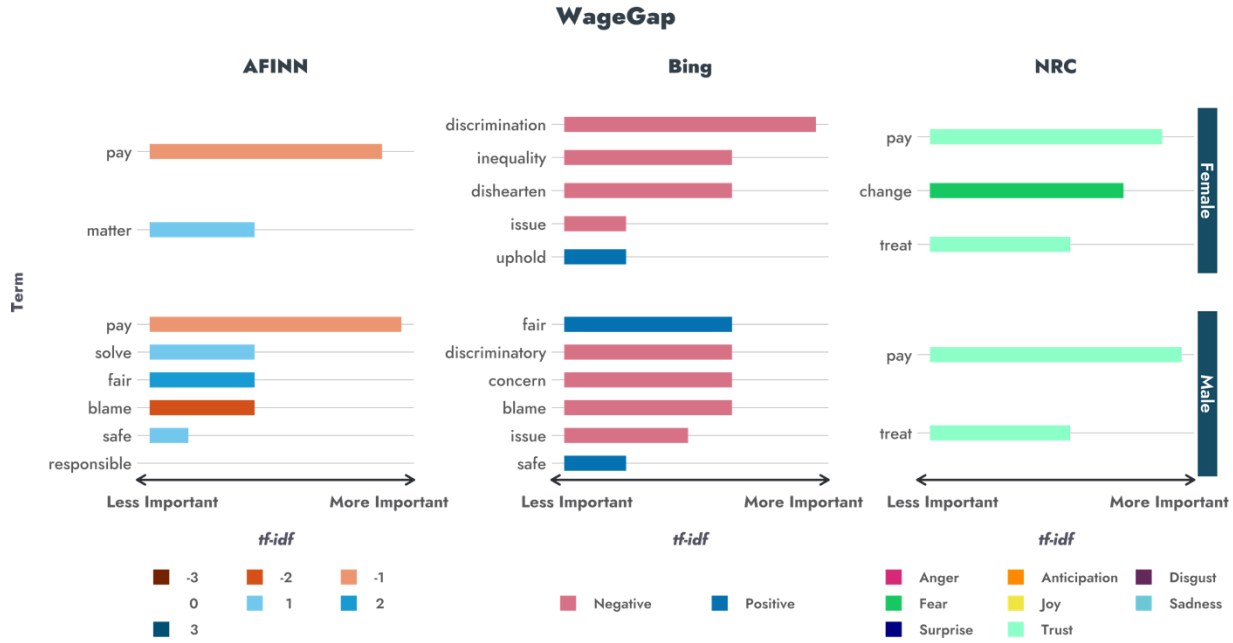

Figure 7. Sample of top 10% of common terms across all Wage Gap activities ranked by *td-idf*. Terms with empty bars indicate a neutral sentiment.

Discussion. The example provides an instance where ML techniques can help an evaluator to quickly process qualitative data when data are plentiful, answers are important but not critical, and the availability of time to make programmatic decisions is limited. In situations where decisions have greater impacts, these results should be viewed as indicators and further quantitative or a separate qualitative study can be used to confirm results. For example, the differences that existed between male and female students and their corresponding responses to the reflection question assessing content learned from the activities were drastic at times and a thematic analysis could have been used to confirm underlying sentiments and provide greater context. The specific activities in which those sentiments were situated enabled the evaluation team context through which to view the student responses.

Without an assessment of the context of a sentiment situated within specific activities, it is difficult to draw reliable recommendations and conclusions about the sentiments. For

example in the algorithmic justice league assignment, while not focused on issues of gender but rather on bias, the term “bias” was in the top 10% of every lexicon for females and elicited very strong emotions, but the term “bias” was altogether absent in the top 10% for male students in each lexicon. And while male and female students responded with different emotions and intensity, the neuroplasticity activity was the most gender-neutral assignment, which was reflected in the results. Finally, male and female students experienced some different emotions on the most gender-centered activity, the wage gap. Male and female students noted varying emotions around pay and negative emotions around discrimination, but female students noted strong negative emotions associating “change”, “inequality”, and “dishearten” with fear, while their male counterparts were more likely to note terms such as “issue”, ”blame”, and “concern”. The differing responses may indicate that male students see the wage gap as a problem but do not take it as personally as the female students.

To change the culture in engineering and computer science, all engineers and computer scientists need to contribute toward the change. The change cannot be motivated only by those in the marginalized groups. While the research team has created multiple activities to change the culture, some tweaks can be made within the activities to more directly ensure all students identify the issues and address them. By having the assignments more explicitly draw on students to provide concrete steps toward mitigating issues, like the one identified in the wage gap activity, the culture is more likely to shift, and non-marginalized students may be more likely to become allies with the marginalized students.

Limitations. The authors acknowledge that a sentiment analysis may not be an acceptable method for all evaluators as there is often a great deal of programming skill

necessary to conduct one properly and efficiently. We also do not dispute that some terms may have been labeled incorrectly because we did not assess each in context and the AI was not trained to understand the reasoning behind one's sentiment, rather only to use existing lexicons in assessing the underlying sentiments associated with the data set. Additionally, the use of unsupervised learning, while beneficial when one does not have data on desired outcomes, lacks the depth of supervised outcomes implying that the AI tagged sentiments without any prior knowledge of the study. This approach only addressed context for terms that had inconsistent sentiment values, thus limiting our scope.

Finally, we collected demographic data from the students on their biological sex, not their gender identity. Thus, our results only apply in the context of sex and may not extend to gender as biological sexes are not synonymous with each other. Future studies should accurately capture gender identities.

Future Research. We presented the use of a single ML technique applied within a specific case under certain circumstances. While this is a tactic that can be understood and applied by many evaluators, additional methods can be used for analysis which may provide better or more informative results. For example, machine learning tools such as hierarchical clustering or topic modeling can be used to uncover likely themes within a body of text whereas neural networks and link predictions in a social network may be used to classify text and impute missing data (Allahyari et al., 2017). These are but a handful of techniques under the broadly defined field of machine learning that can provide further information on the sentiments we have already discovered and may even serve as a check in certain situations. As the current

data set grows, we anticipate implementing many of these in addition to using variants of the current sentiment analysis both at the sentence and paragraph levels.

Methodological Benefits

The true advantage of using a sentiment analysis lies in an evaluator's ability to use a quantitative approach in analyzing text, especially in those circumstances where the breadth of a data set or constraints associated with resources such as time and limits on the number of personnel with qualitative expertise makes it very difficult for practitioners to perform a comprehensive analysis. Additionally, sentiments can be presented graphically providing an evaluator with numerous approaches to visualizing both the data and results. Finally, since the method is purely quantitative, comparing sentiments across and between groups can be accomplished easily.

From a practitioner standpoint, results can be used to identify underlying attitudes associated with open-ended response data or to assess the feelings of groups of people without directly asking, both that can be used for formative or summative purposes. Furthermore, evaluators can use a sentiment analysis to discover key aspects of a program and its associated activities that stakeholders or sponsors care about while addressing the principal concerns and goals of the program participants. While still a maturing tactic, assessing the efficacy of tasks through a sentiment analysis is a promising method for uncovering themes, the connectedness of student responses, and determining what areas of a program merit investigations.

Bibliography

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*

Allan, E. J., & Madden, M. (2006). Chilly Classrooms for Female Undergraduate Students: A Question of Method? *The Journal of Higher Education*, *77*(4), 684-711.

Årup Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv e-prints*. Retrieved March 01, 2011, from https://arxiv.org/abs/1103.2903

Bates, M. (1995). Models of natural language understanding. *Proceedings of the National Academy of Sciences*, *92*(22), 9977-9982. https://doi.org/10.1073/pnas.92.22.9977

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Christina, M. V., Hocevar, D., & Hagedorn, L. S. (2007). A Social Cognitive Construct Validation: Determining Women's and Men's Success in Engineering Programs. *The Journal of Higher Education*, *78*(3), 337-364.

Corbett, C., & Hill, C. (2015). Solving the Equation: The Variables for Women's Success in Engineering and Computing.

Ding, X., Liu, B., & Yu, P. S. (2008). *A holistic lexicon-based approach to opinion mining* Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, California, USA.

Feinerer, I., & Hornik, K. (2018). *tm: Text Mining Package*. In https://CRAN.R-project.org/package=tm

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, *25*(5), 1-54. https://doi.org/10.18637/jss.v025.i05

Grefenstette, G., & Tapanainen, P. (1994). What is a word, What is a sentence? Problems of Tokenization. The 3rd International Conference on Computational Lexicography (COMPLEX'94), Budapest.

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*(3), 1-25.

Hall, R. M., & Sandler, B. R. (1982). *The classroom climate : A chilly one for women?* Project on the Status and Education of Women, Association of American Colleges.

Holloway, I., & Todres, L. (2003). The Status of Method: Flexibility, Consistency and Coherence. *3*(3), 345-357. https://doi.org/10.1177/1468794103033004

Hughes, R. (2012). GENDER CONCEPTION AND THE CHILLY ROAD TO FEMALE UNDERGRADUATES' PERSISTENCE IN SCIENCE AND ENGINEERING FIELDS. *18*(3), 215-234. https://doi.org/10.1615/JWomenMinorScienEng.2013003752

Karlsson, F. (1994). *Yleinen kielitiede [General Linguistics]*.

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *44*(4), 491-511. https://doi.org/10.1177/0165551517703514

Kilicoglu, H., Peng, Z., Tafreshi, S., Tran, T., Rosemblat, G., & Schneider, J. (2019). Confirm or refute?: A comparative study on citation sentiment classification in clinical research publications. *Journal of Biomedical Informatics*, *91*, 103123. https://doi.org/https://doi.org/10.1016/j.jbi.2019.103123

Könings, K. D., Brand-Gruwel, S., & van Merriënboer, J. J. G. J. I. S. (2011). Participatory instructional redesign by students and teachers in secondary education: effects on perceptions of instruction [journal article]. *39*(5), 737-762. https://doi.org/10.1007/s11251-010-9152-3

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis : a methods sourcebook* (Third edition. ed.). SAGE Publications, Inc.

Mohammad, S. M., & Turney, P. D. (2013a). Crowdsourcing a Word-Emotion Association Lexicon. *arXiv e-prints*. Retrieved August 01, 2013, from https://ui.adsabs.harvard.edu/\#abs/2013arXiv1308.6297M

Mohammad, S. M., & Turney, P. D. (2013b). CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON. *29*(3), 436-465. https://doi.org/doi:10.1111/j.1467-8640.2012.00460.x

Mullen, L. A., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software*, *3*, 655.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing. http://www.R-project.org/

Rajaraman, A., & Ullman, J. D. (2012). *Mining of massive datasets*. Cambridge University Press.

Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., & Dodds, P. S. J. E. D. S. (2017). Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs [journal article]. *6*(1), 28. https://doi.org/10.1140/epjds/s13688-017-0121-9

Rinker, T. W. (2018). *textstem: Tools for stemming and lemmatizing text*. In http://github.com/trinker/textstem

Roy, J. (2018). Engineering by the numbers. In *2018 ASEE Profiles of Engineering and Engineering Technology Colleges* (Vol. 2018, pp. 524). American Society for Engineering Education.

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, *1*(3).

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach* (1st ed.). O'Reilly Media.

Trapani, J., & Hale, K. (2019). *Higher Education in Science and Engineering. Science and Engineering Indicators 2020* (NSB-2019-7). National Science Foundation.

Wang, S., & Manning, C. D. (2012). *Baselines and bigrams: simple, good sentiment and topic classification* Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, Jeju Island, Korea.

Weissman, G. E., Ungar, L. H., Harhay, M. O., Courtright, K. R., & Halpern, S. D. (2019). Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of Biomedical Informatics*, *89*, 114-121. https://doi.org/https://doi.org/10.1016/j.jbi.2018.12.001

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. In https://CRAN.R-project.org/package=tidyverse

Zhang, L., & Liu, B. (2017). Sentiment Analysis and Opinion Mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1152-1161). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_907