CREATING AND TESTING SPECIALIZED DICTIONARIES FOR TEXT ANALYSIS

Roman Taraban
roman.taraban@ttu.edu
Jessica Pittman
jessica.pittman@ttu.edu
Taleen Nalabandian
taleen.nalabandian@ttu.edu
Winson Fu Zun Yang
winson.yang@ttu.edu
William M. Marcy
william.marcy@ttu.edu
Texas Tech University, Lubbock, USA
Srivinasa Murthy Gunturu
srinivasamurthygunturu@gmail.com
Tata Consultancy Services, Chennai, India

Received June 12, 2019; Revised June 26, 2019; Accepted June 28, 2019

Abstract. Practitioners in many domains–e.g., clinical psychologists, college instructors, researchers–collect written responses from clients. A well-developed method that has been applied to texts from sources like these is the computer application Linguistic Inquiry and Word Count (LIWC). LIWC uses the words in texts as cues to a person's thought processes, emotional states, intentions, and motivations. In the present study, we adopt analytic principles from LIWC and develop and test an alternative method of text analysis using naïve Bayes methods. We further show how output from the naïve Bayes analysis can be used for mark up of student work in order to provide immediate, constructive feedback to students and instructors.

Keywords: *text analysis, machine learning, LIWC, naïve Bayes*.

Тарабань Роман, Піттман Джесіка, Налабандян Талін, Янґ Вінсон Фу Зун, Марсі Вільям, Ґунтуру Шрівінаса Мерті. Створення та тестування спеціалізованих словників для аналізу тексту.

Анотація. Робота фахівців-практиків у багатьох галузях, наприклад, клінічних психологів, викладачів коледжів, дослідників передбачає збір письмових відповідей їхніх клієнтів чи студентів. Добре розроблений метод, який застосовується сьогодні до текстів такого типу, − це комп'ютерний додаток Linguistic Inquiry and Word Count (LIWC). Програма LIWC трактує слова в текстах як індикатори ментальних процесів людини, її емоційних станів, намірів і мотивів. У статті використано аналітичні принципи LIWC, розроблено та протестовано альтернативний метод аналізу тексту з використанням методів наївного баєсового класифікатора. Автори демонструють, як результати аналізу за наївним баєсовим класифікатором можуть бути використані для аналізу студентської роботи з метою надання негайного, конструктивного зворотного зв'язку і студентам і викладачам.

Ключові слова *аналіз тексту, машинне навчання, LIWC, наївний баєсів класифікатор*.

East European Journal of Psycholinguistics, 6(1), 65–75. https://doi.org/10.5281/zenodo.3377557

1. Introduction

The linguist, Edward Sapir, believed that "language and our thought-grooves are inextricably interwoven, [and] are, in a sense, one and the same" (in Salzmann, 2004, p. 43). An assumption that characterizes contemporary thinking across many domains of research and applications is that the language a person uses can reveal a great deal about that person, including thoughts, feelings, motivations, and personality. Pennebaker and King (1999) proposed that "the way people talk about things reveals important information about them" (p. 1297). Elsewhere, Tausczik and Pennebaker (2010) suggested that "The words we use in daily life reflect who we are and the social relationships we are in" (p. 25).

Practical applications of language analysis can be traced as far back as the psychoanalytic work of Freud (Tausczik & Pennebaker, 2010). Whereas early work was slow and tedious, recent advances in technology have enabled analyses of large language samples from sources like product reviews, forums, blogs, social networks, and mental health settings. These analyses have been used productively to achieve a variety of goals, for example, in business for sentiment analysis, and in clinical settings to treat depression. The focus of the work presented in this paper is on using machine tools to analyze the semantic content of college students' written class work and to provide automated feedback regarding the quality and coverage of their responses for specific writing tasks. The analytic procedure that we describe can be applied to a wide variety of data and is not limited to the college course data we present here.

A successful and widely applied machine tool for text analysis is Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC uses pre-defined dictionaries as the basis of its computing power. In our approach to text analysis, we adopt the basic assumption underlying LIWC, which is that words in a text can function as statistical variables and thereby provide the basis for quantitative analysis. We further assume that words, as variables in an analysis, have a weighted relationship to the message that is being conveyed. That is, some words are more important than others. We use naïve Bayes analysis to incorporate these ideas into a general method for constructing dictionaries for specific language corpora of interest to practitioners.

1.1. Machine Tools for Text Analysis

Machine tools for analyzing the content of language samples are based on the general assumption that aspects of the semantic structure of text can be recovered through algorithmic methods. The approaches across machine methods vary, with some systems constructing high-dimensional semantic spaces of correlated words (Landauer, Foltz, & Laham, 1998; Lund & Burgess, 1996), others relying on pre-defined words to identify cognitive and affective categories (Pennebaker et al., 2015), and yet others searching out topics across samples of documents based on distributions of words within and across the documents (Blei, Ng, & Jordan, 2003). In the next two subsections, we briefly describe LIWC, which provides a framework within which to understand the methods that we develop, and naïve

Bayes analysis, which is a well-known algorithm for calculating complex conditional probabilities.

1.2. LIWC

Highly selective lists that define categories are the heart of the LIWC program. These lists were developed over the course of decades and in consideration of extensive samples of texts (Tausczik & Pennebaker, 2010). When presented with a text for analysis, the LIWC program searches through the text, word by word, and compares each word with those in the pre-defined categories of words. The percentage of words in each LIWC category–as determined by the presence of words that define that category–are subsequently calculated. The LIWC program reports the percent of words for 125 categories with nearly 6,400 words or word stems (Pennebaker et al., 2015) defining these categories. Examples of LIWC categories include those that are:

- Content-oriented: work (*hire*, *review*, *memo*) and home (*laundry*, *backyard*, *family*)
- Grammatical: articles (*a*, *an*, *the*) and prepositions (*over*, *under*, *between*)
- Psychological: positive (*pleasant*, *hopeful*, *compassion*) and negative (*jealousy*, *loneliness*, *terrified*) emotion (Pennebaker et al., 2015).

Overall, LIWC uses the words in texts as cues to a person's thought processes, emotional states, intentions, and motivations. The influence of LIWC on text analysis has been broad, with translations of LIWC dictionaries into Catalan (Massó, Lambert, Penagos, & Saurí, 2013) and Dutch (Boot, Zijlstra, & Geenen, 2017; Van Wissen, & Boot, 2017), among other languages.

A limitation of the LIWC approach is the reliance on pre-defined dictionaries and categories for classification. Specifically, the dictionaries are constructed to identify and quantify specific categories. Dictionaries of grammatical categories (e.g., first-person plural pronouns) that stem from the English language are face-valid, whereas dictionaries of psychological (e.g., cognitive processes) or content-oriented (e.g., family) language categories were constructed by judges and, thus, are more likely to represent inaccurate categorization (Newman, Groom, Handelman, Pennebaker, 2008; Pennebaker et al., 2015). Applicable to the current study, the dictionaries are susceptible to missing relevant categorical information in a target set of essays because the categories of interest may not be well represented by the LIWC categories. Our method, using naïve Bayesian methods, attempts to bypass this limitation.

1.3. Naïve Bayes

Naïve Bayes is an algorithm based on the calculation of conditional probabilities. The basic equation for estimating the probability of some category X given some variable y is:

$$P(X \mid y) = P(X \cap y) / P(y) \qquad (1)$$

The probability of X AND y, $P(X \cap y)$, in this equation is equal to P(y | X) * P(X). If we treat the words in a text as variables, then there are multiple predictors, y_i. Naïve Bayes treats each of the words as independent predictors, so that the

numerator in (1) becomes:

$$P(y_1 \mid X) * P(y_2 \mid X) * P(y_3 \mid X) \ldots P(y_i \mid X) * P(X)$$

The simplifying assumption of independence of predictors–i.e., there is no consideration of interactions between predictors–allows the algorithm to easily compute probable classifications based on large numbers of predictors.

Naïve Bayes can be used to build classifiers using supervised learning methods. Basically, for some set of data, human raters classify the instances. The naïve Bayes classifier computes the strongest predictors within those instances in order to best match human classifications, and can then apply these predictors to new instances. In the present case, the predictors are the words in students' essays. The significant difference between LIWC and a naïve Bayes classifier, which we want to emphasize here, is that LIWC applies pre-defined and fixed predictors–i.e., the words that define the LIWC categories–in order to identify predefined categories. Through naïve Bayes, the researcher defines the categories of interest for a sample of texts and naïve Bayes discovers the predictors in the sample that are most strongly associated with those classifications. Thus, naïve Bayes is able to create specialized dictionaries tailored to the needs and interests of the researcher and in consideration of the available texts.

The present work is largely exploratory. We have several related goals:

- To build a naïve Bayes classifier to identify specific content in students' open-ended essays
- To test the classifier's ability to reliably transfer its knowledge to new essays
- To identify the most reliable predictors that naïve Bayes uses in its classifications
- To mark up students responses as a form of constructive feedback to students.

2. Methods

A website was created entitled Ethical Engineer https://EthicalEngineer.ttu.edu (see Figure 1) to allow engineering students to read and respond to case studies posing ethical issues that have arisen in engineering. The website outlines a prompt for students to respond to, posing an array of questions to consider when approaching these ethical issues. The Ethical Engineer currently includes three case studies. The present study analyzed student comments to one of three case studies that are presented on the website. The case study is titled "Which is More Important – Environmental Concern or Economic Growth?" and can be read in full on the website.

Participants were primarily students enrolled in an undergraduate ethics course offered at Texas Tech University, as well as students from participating institutions abroad, primarily India. Students participated on a volunteer basis.

2.1. Procedure

The 119 independent comments to the case study that were available on the website at the time of this study were selected for analysis. The instructions for

submitting a comment are shown in Figure 2 (color coding was not used on the website and is used in the figure to highlight the categories that were analyzed using naïve Bayes).[1] An example of a typical student comment is shown in Figure 3.

Figure 1. Screen Shot of https://EthicalEngineer.ttu.edu Case Study Page

Submit a Comment
As you read and analyze case studies your reflective comments are invited on some or all of the following. a) As part of your analysis include information on the stakeholders and how they are impacted both positively and negatively.
b) What knowledge and skills are needed to implement sophisticated, appropriate and workable solutions to the complex global problems facing the world today?
c) What interdisciplinary perspectives would help identify innovative and non-obvious solutions?
d) What insights can you articulate, based your culture and other cultures with which you are familiar, to help understand your worldview and enable greater civic engagement?
e) What is your position on the right thing(s) to do?

Figure 2. Instructions for Submitting a Comment to a Case Study on the Website https://EthicalEngineer.ttu.edu (Color coding and lower-case lettering are used here to indicate the categories of interest to the reader – details below)

[1] Color coding in this and the other examples in this paper is visible in the online version of the paper but not in the paper copy of the Journal.

petroleum413 April 9, 2019 at 12:27pm

In the case study "Which is more important- Environmental Concern or Economic Growth" by Dr. Majumdar, the situation examined is about an area in India known as Trombay economic growth and pollution due to big oil companies. Trombay and the surrounding areas economies began to expand rapidly due to the big oil companies drilling and refineries, but with this expanding company came many negative consequences. The environment and the surrounding communities were greatly affected by the pollution which was being created by the drilling sites and refineries. One way to help prevent these situations from occurring is for engineers and large oil companies to know the most effective drilling and refinement process which minimize negative environmental impact. Secondly, problem solving skills and the ability to communicate respectively to people of other cultures are an essential tool to solving the complex global problems created by big oil companies. Also, knowledge of safe disposal practices is an essential tool to solving the difficulties facing the world today. Third, by engineers having interdisciplinary perspectives such as knowledge about chemistry and economics would assist in detecting innovative, non-obvious solutions to balancing economic growth and the impact on the environment. Fourth, civic engagement is an essential device to understanding the balance between environmental concern and economic growth. In many cultures certain land is considered sacred, holy, or historical significance. In the event that there is holy, sacred, or historical lands is near drilling sites, engineers with knowledge about the locations of these land can enable superior civic engagement. Lastly, the balance between economic growth and environmental concern is an extensive ethical concern. I believe engineers should take precautions to prevent negative environmental impact. A more expensive piece of equipment may affect the company's profits but will eliminate potential problems in drilling or refinement is worth the expense. Also, I believe countries who do not have strict environmental regulations should not be taken advantage due to less restrictive laws.

Figure 3. Sample Comment to the Case Study "Which is More Important…" Displayed on the Website https://EthicalEngineer.ttu.edu

2.2. Classification by Human Raters

Each comment was parsed into sentences. The 119 comments resulted in 1631 sentences. Two trained researchers carried out classification of each sentence as pertaining to one of five possible categories, as indicated by the color coding in Figure 2: a) Stakeholder, b) Knowledge and Skills, c) Interdisciplinary Perspectives, d) Cultural Understanding, e) Right Action. A sixth category, Other, was used to classify sentences that did not fit one of these categories. For classification, the researchers independently classified the sentences. The researchers then reviewed their combined classifications and resolved cases of disagreement through discussion.

The 1631 sentences were divided into a training set, based on 84 student comments consisting of 1196 sentences, and a test set based on 35 student comments consisting of 435 sentences. Inter-rater agreement was similar for the training set (77 % agreement) and test set (75 % agreement).

2.3. Naïve Bayes Classifier

The 1196 training sentences and human classifications were input into a comma separated values (.csv) file, a portion of which is shown as an example in Figure 4.

Comment Sentences	Classification
This interdisciplinary perspective always should be applied when people make judgements about any solutions.	Interdisciplinary
In the case study above, the stakeholders indeed make a lot of money of drilling oil in the farm, but meanwhile they sacrifice many other residents' lives and health around this area.	Stakeholder
The corporations around the world have been more and more active and tightly interactive.	Culture
People would be more open and have higher tolerance of culture difference.	Culture
However, there are still some taboos that should be considered before stakeholders make any decisions.	Right
Good research and understanding from both sides are the prerequisites for good cooperation.	Interdisciplinary

Figure 4. Example of Portion of Comma Separated Values (.csv) File Used for Training and Testing a Naïve Bayes Classifier

The .csv file constituted the input to naïve Bayes, which was implemented in R https://cran.r-project.org/web/packages/naivebayes/naivebayes.pdf through R-Studio, using package e1071 and Laplace smoothing. Numbers, stop words (e.g., function words like *the, if, on*), and punctuations were removed, and stemming (e.g., reducing *trouble, troubles, troubling* to *troubl*) was applied. The modification of the data, as described, resulted in 584 word stems across the 1631 sentences. Therefore, the naïve Bayes analysis was based on 584 predictors. These predictors were tested against each sentence, which allowed naïve Bayes to estimate the most likely classification for each sentence.

3. Results

A confusion matrix showing frequencies (percent of total sentences shown below frequencies) of agreements and disagreements between human raters and the naïve Bayes classifier for 435 new sentences is shown in Figure 5. Matches between human raters and the naïve Bayes classifier indicate a modest 61.1 % level of agreement between humans and machine. Given that the two human raters initially agreed on classifications 75 % of the time for these test data, we should not expect the Bayes classifier to agree with the human raters at a 100% level. Instead, it makes more sense to think of the Bayes classifier as another rater, in which case there is a 14 % discrepancy between the ability of humans to make the classifications and the machines ability to make comparable classifications.

3.1. Text Markup for Feedback

Output from naïve Bayes was used to mark up students' comments as potential feedback to students and instructors. The visual displays showing mark up through color coding and figures (See Figures 6 and 7 below) are implemented using the Shiny application http://shiny.rstudio.com/ in R Studio.

One form of text markup is to use the most probable naïve Bayes classification of each sentence (Culture, Interdisciplinary, etc.) to mark up a student's submission through color coding. This form of markup can provide students and instructors with immediate visual feedback regarding coverage of the recommended points to address, as indicated in the instructions for leaving comments. The markup also shows the distribution of comments to the classifications targeted in the naïve Bayes analysis. The color coding of sentences in a student's comment is supplemented by the Shiny application with a bar graph and radar graph (See Figure 6), providing students with additional information about their coverage of the points targeted in the instructions for leaving comments.

	Actual(Human Raters)						
Predicted	Culture	Interdisciplinary	Other	Right	Skills	Stakeholder	Row Total
Culture	23 0.053	2 0.005	4 0.009	4 0.009	2 0.005	1 0.002	36
Interdisciplinary	0 0.000	28 0.065	0 0.000	1 0.002	4 0.009	0 0.000	33
Other	6 0.014	2 0.005	13 0.030	2 0.005	5 0.012	17 0.039	45
Right	8 0.018	12 0.028	23 0.053	69 0.159	13 0.030	26 0.060	151
Skills	1 0.002	3 0.007	0 0.000	4 0.009	25 0.058	1 0.002	34
Stakeholder	5 0.012	2 0.005	7 0.016	9 0.021	5 0.012	107 0.247	135
Column Total	43	49	47	89	54	152	434

Figure 5. Confusion Matrix for New Classifications

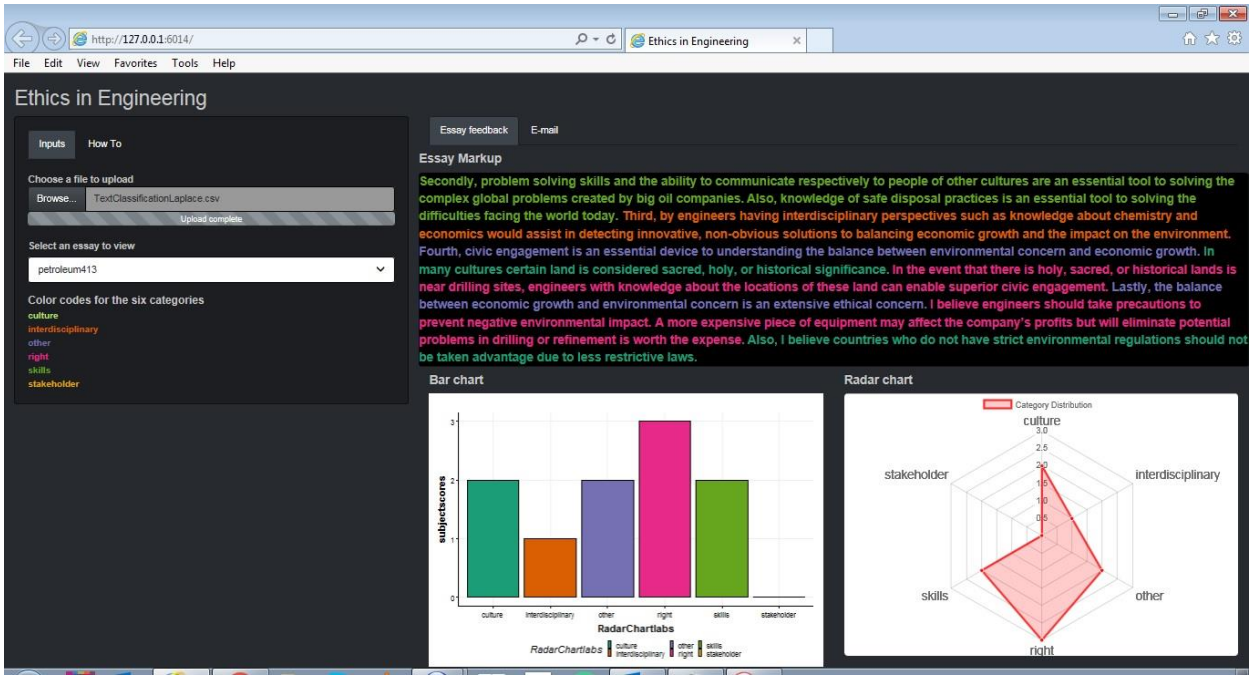

Figure 6. Example of Feedback Showing Most Likely Classifications of Sentences Based on Naïve Bayes Output and Mark Up Using Shiny App

A second form of markup, to provide another form of feedback to the user and the instructor, can be carried out as follows. First, calculate simple Bayesian probabilities for each predictor (stem) for each category (stakeholder, inter-disciplinary…), next rank order the predictors for each category, and, finally, use a

subset of ranked predictors (e.g., top 10) for each category in order to mark up the text. This feedback can be used to make more explicit the strongest conceptual elements within an essay. An example of text markup using this method is shown in Figure 7.

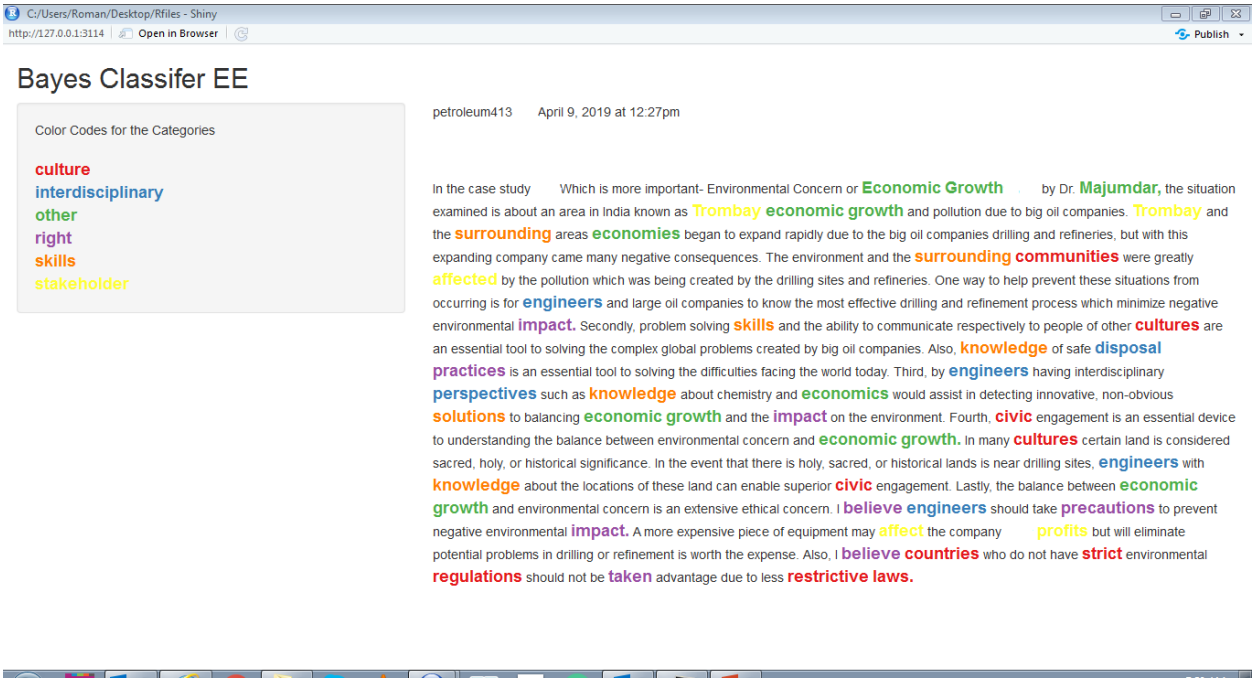

Figure 7. Example of Feedback Using Bayesian Probabilities of Most Reliable Stems and Markup Using Shiny App

4. Conclusions

The present analysis of a sample of students' comments to an engineering case study using naïve Bayes showed fair agreement between machine and human classification. We believe the algorithm will come into closer agreement with humans as we increase the amount of data for training. Further tests will show whether this is indeed the case.

If our approach to constructing specialized dictionaries using Bayesian analyses– in lieu of pre-defined dictionaries as employed by language software, such as LIWC–proves successful, several potential benefits emerge for instructors. First, content analysis can be tailored to students' vocabulary levels, regional vernacular, and other word choice factors. Second, the method provides for a flexible range of analysis, i.e., it affords the analysis of short responses or longer essays. Finally, the method allows instructors to focus on course-related subject matter, i.e., classifiers can be directed to specific course topics. Overall, the very practical benefits of the Bayesian methods we describe are an ability to quickly bring a classifier up to speed, to continually update the classifier with additional human assessments, to tailor the classifier to the specific needs and goals of an instructor, and to merge naïve Bayes code with other code necessary for creating an interface for student input and feedback.

The current goal of this project is to classify comments from the Ethical Engineer website according to topics (e.g., Stakeholder, Interdisciplinary…) of current interest to the instructor. However, we envision further extensions of this work. For example, it is possible to add to the classifications that naïve Bayes identifies, for instance, classifying sentences in student submissions, or other sources, according to binary classifications like

- descriptive/analytic
- productive/unproductive
- high/low quality.

These classifications could be supplemented by identifying the strongest predictors that naïve Bayes used to make those classifications and marking them up in the submission, as in Figure 7.

The methods described here are not without limitations. The algorithm treats predictors as independent, which is handy statistically and from the perspective of cognitive modeling, but which also introduces a limiting heuristic. That is to say, conceiving of the classification process as a compilation of independent predictors ignores interactions between predictors. Knowledge of how predictors interact and combine into more complex constructions like propositions (Kintsch, 1998) would significantly extend the analysis and feedback that could be provided to students and instructors.

Finally, although the naïve Bayes and markup methods we describe here can provide useful feedback to students, effective feedback may still require human judgment to provide students with input on the depth, insights, empathy, and creativity of their responses. These are human- and machine-processing questions and challenges that still remain.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, *6*(1), 65-76.

Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality, 42*(1), 96-132.

Hsieh, H-F., & Shannon, S. E. (2005).Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 277-1288.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. New York: Cambridge University Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.

Massó, G., Lambert, P., Penagos, C. R., & Saurí, R. (2013). Generating New LIWC Dictionaries by Triangulation. In: *Asia Information Retrieval Symposium* (pp. 263-271). Springer, Berlin, Heidelberg.

Newman, M., Groom, C.J., Handelman, L.D., & Pennebaker, J.W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45(3)*, 211-236.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC 2015*. Austin, TX: University of Texas at Austin.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24-54.

Van Wissen, L., & Boot, P. (2017). An electronic translation of the LIWC dictionary into Dutch. *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference.* (703-715).