Evaluating the Teaching Evaluations of One Hundred North American Schools

Dr. Haroon Malik, Marshall University

Dr. Malik is an Associate Professor at the Department of Computer Sciences and Electrical Engineering, Marshall University, WV, USA.

Dr. Wael A. Zatar, Marshall University

Dr. Zatar serves as the Dean of the College of Information Technology and Engineering at Marshall University.

©American Society for Engineering Education, 2020

Evaluating the Teaching Evaluations of One Hundreed North American Schools

I. INTRODUCTION

For service providing establishments, collecting the perception of the quality of service received by the customer has become a norm. This includes organizations providing education. Instruments used in educational establishments solicit student data of two types: (1) on a predefined rubric, aspects of quality, where the respondents mark-off their perception on a predefined, Likert-like scale. (2) the respondents write their very own views on concepts of quality, and the scale used for the assertion. This is wholly unstructured data. While the first type provides a baseline, which facilitates comparisons, the second set provides the only true measure of the quality of service as perceived by that receiver and delivered by that provider. The authors believe that the portion of data bereft of scale and rubric, collected in course evaluations, is perhaps the most useful for the individual instructor's professional development. The main **objective** of the work-in-progress (WIP) is to develop a methodology to: (a) automatically extract assertions of perceived quality of teaching using machine learning techniques. (b) provide a mechanism to compare instructors based on the extracted assertion/qualities. The **contributions** of the paper are (a) methodology to mine teaching evaluation and (b) an open-source tool to facilitate educational establishments execute empirical studies and students perform exploratory analytics on the teaching evaluations. The tool supports a wide variety of data formats, does not require any domain knowledge for its operation and will be released under a general public license (GPL).

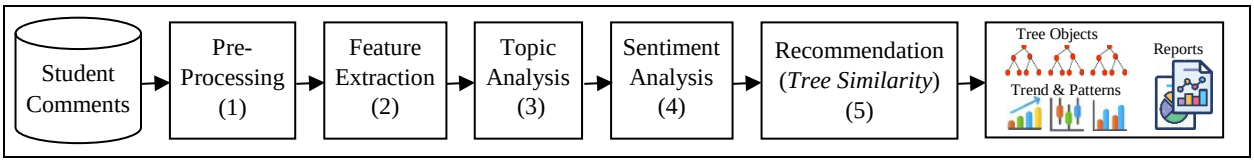

Figure 1. A high-level overview of the proposed methodology

II. METHODOLOGY

Figure 1 shows an overview of the proposed methodology and consist of the following steps:

Pre-Processing: The proposed feature extraction algorithm cannot be directly applied to the student comments from most of the readily available sources. This is due to the reason: (a) a large portion of reviews are submitted from devices such as phones and tablets with no full keyboard (b) students use texting slang words, jargons and abbreviations (i.e., BTW-by the way, H8-hate, W@-what). Therefore, the following pre-processing steps are performed. **(1)**. *Noun, verb and adjective extraction* — The part of speech (POS) tagging functionality of the Natural Language Toolkit, NLTK4, is used for identifying and extracting the nouns, verbs, and adjectives in the student reviews. **(2)**. *Stop-word removal* — Stop-words in student comments are removed to eliminate terms that are very common in the English language (e.g., "and", "this", and "is"'). The standard list of stop-words provided by 'Lucene5' is used. The list is also expanded to include words that are common in user reviews but are not used to describe features. **(3).** *Lemmatization* —Wordnet lemmatizer [1] from NLTK is used for grouping the different inflected forms of words with the same part of speech tag which are syntactically different but semantically equal. For example, the terms describing the verbs "sees" and "saw" are grouped into the term "see".

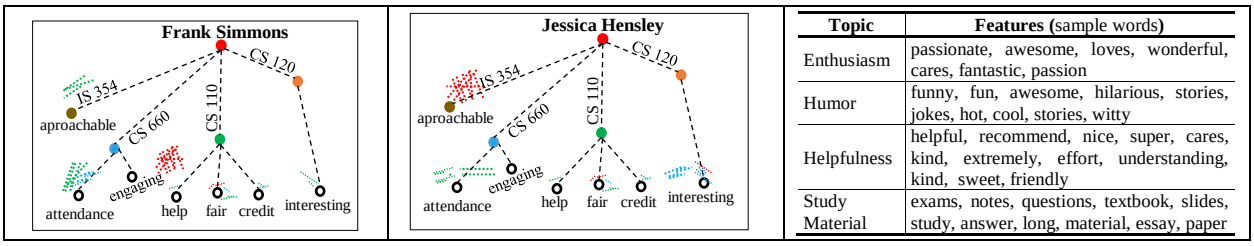

Figure 2. Sentiment Analysis of hot features of two profs (Green dots - positive; Red dots - negative; Blue dots , neutral opinion

Topic	Features (sample words)
Enthusiasm	passionate, awesome, loves, wonderful, cares, fantastic, passion
Humor	funny, fun, awesome, hilarious, stories, jokes, hot, cool, stories, witty
Helpfulness	helpful, recommend, nice, super, cares, kind, extremely, effort, understanding, kind, sweet, friendly
Study Material	exams, notes, questions, textbook, slides, study, answer, long, material, essay, paper

Table1. A few topics and their corresponding features

Feature Extraction: The foremost and the key step of the methodology is feature extraction. The task of feature extraction in this paper is to transform student comment data into a feature space that can best describe the interests of students who comment on the instructors or their associated courses. More specifically, to extract only the relevant instructor/course features that appeared in the comments. In a data-mining task, all the features are generally regarded as nouns. Hu and Liu [2] also treated frequent nouns and noun-phrases as product feature candidates, i.e., opinion targets. Similarly, Gupta et al. [3] also testified that all features are nouns/noun-phrases. Therefore, a straight forward way to extract the features from the student's evaluation/review is to scrape all the nouns, from each sentence in the review, from the entire dataset (as bag-of-words). As a pilot study, 600 reviews consisting of 2,452 lines from "*Ratemyprofessor.com*" website (RMP) are annotated. The default list of features provided by RMP (as tags) is compared with that of all nouns collected as bag-of-words. It is found that bag-of-words do contain 98% of the features provided by RMP. This means that even this naïve approach of simply part-of-speech tagging the entire review set and separating the noun is useful in terms of precision.

Topic Modeling: Many of the extracted features, though syntactically different, inherently share a common theme, i.e., topic (Table 1). To group features into relevant topics, a popular topic modeling technique called latent Dirichlet allocation (LDA) is used, as it is best suited for the research goal of finding discussion topics in natural language text documents, i.e., student reviews. LDA is a statistical topic modeling technique, which means that LDA represents topics as probability distributions over the words in the corpus, and it represents documents (i.e., reviews) as probability distributions over the discovered topics. LDA creates topics when it finds sets of features that tend to co-occur frequently in the documents of the corpus. Often, the features in a discovered topic are semantically related, which gives meaning to the topic as a whole. Table 1 lists some of the topics extracted from RMP data along with their corresponding features.

Sentiment Analysis — Sentiment analysis is used as a part of the methodology to explore student's opinions about a feature. In general, the opinions of students can be classified into three categories; positive, negative and natural. Most of the time, students use predictable words while giving comments to express perceived experiences with an instructor. At times, there are many slang words used by students such as '*phew*', '*huhh*', '*oh man*' and '*Ummmm*' that act as noise and don't contribute towards sentiment analysis. A sentiment word bank is created by removing such seldom-used words with the help of WordNet [1]. Afterward, a bit map is established by tagging the words in the comments that appear in the word-bank with a value of '1' and others with '0'. The bit map facilitates the next step of the methodology in displaying the sentiment orientation of

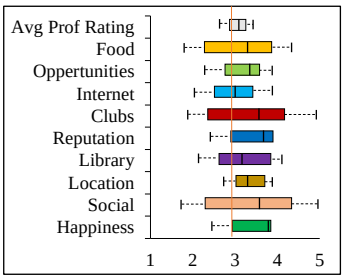

Figure 3. The distribution of the quantitative score of the top-50 universities

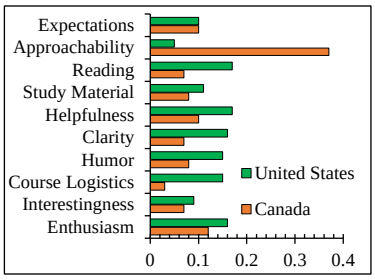

Figure 4. Topic distribution among good professors

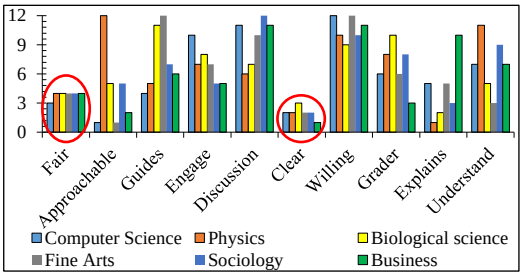

Figure 5. Top-10 topics across different groups

the features (like positive, negative, or neutral) in each sentence of student review, further centered on a topic as shown in figure 2.

Recommendation — The last step of the methodology is aimed to expedite exploratory analysis by outputting (a) list of features and topics relevant to an instructor, (b) topics evolution graphs and unfolding of corresponding sentiments distributions over a period of time by replying the instructor evaluation history and (c) comparable tree-like structure of instructor courses with most discussed topics, i.e., hot topics (figure 2). Due to space constraints, the paper provides a high-level overview of the working of the algorithm for discovering the concept hierarchy of instructor's most discussed topics for a course and producing an output as a hierarchical tree structure. The algorithm scans through the text, i.e., student comments, once and finds out the hierarchy of the high term frequency-inverse document frequency (TFIDF) words. Suppose there are two random feature words in instructor's reviews: W_1, W_2. The reviews set that contains all the appearance of word W_1 is namely set C_1. Similarly, the reviews set that contain all the appearance of word W_2 is namely set C_1. If set C_1 is a superset of set C_2, then more likely, W_2 is a sub-concept of W_1. A tree structure is used to express the hierarchy like W_1, W_2 instead of a lattice, or a table to ease making a visual comparison among instructors evaluations.

III. ANALYSIS AND INTERMEDIATE RESULTS

The WIP uses two data sources, i.e., RateMyProfessor.com (RMP) and the author's university course evaluations (AUCE), to evaluate the proposed methodology. The AUCE dataset is small, containing two thousand and twenty-two reviews from all the engineering students over a period of five years. The small data-set size facilitates manual validation of the features extracted, topics uncovered and sentiment tagging. Whereas the RMP dataset was used on account of its multiple endorsing characteristics: (a) most academics are familiar with the dataset. (b) students from over 7,000 universities voluntarily provide quantitative and verbal reviews, (c) the publically available data set allows other researchers to replicate our methodology and compare it with that of their own. The collected data from RMP includes student's verbal comments, RMP's pre-declared tags (e.g., lots of homework, inspiration, tough grader) as well as quantitative scores for descriptive factors like location, food, reputations, opportunity, facilities, and average professor ratings, besides the university name, city, state, and number of faculty/instructor information.

The WIP reports high-level findings of understanding teaching evaluations for hundred top universities, i.e., fifty USA and fifty Canadian universities. Noteworthy findings are **(a)** *Whereas good instructors had at least 92% good reviews the others (i.e., bad, as perceived by students)*

continue to accumulate bad reviews. The finding is consistent across the two data sets. **(b)** *Students in Canada seem to be more concerned with Approachability and Study Materials, whereas students from U.S. universities appear to talk about 'Reading/Discussions' and 'Clarity'* (Figure 4). The finding was clinched by using the data corresponding to "good professors", as perceived by students, extracting top-10 topics from their teaching evaluations and looking into the topics distributions (the topic distribution for each group, i.e., USA and Canada, in Figure 4 add up to 100%). **(c)** *There is indeed a difference between what is considered to a 'good instructor' by different groups* (Figure 5). *Despite*, there is some consensus on 'Clear' and 'Fair' to be an essential feature of a good instructor among the groups and is evident from Figure 6. (**d)** *For universities with large number of 'good instructors', all the descriptive factors (listed on the RMP) are rated above average (as expected). Happiness is the highest-rated factor and average professor rating being the lowest-rated factor (surprisingly)* (Figure 3). The finding was concluded by the three-step method. First, for each of the institutions, instructors who maintained a very positive sentiment polarity on student's reviews (90% and above) are selected. Second, creating an index by ranking the institutions based on the computed ratio between selected instructors, and the total number of instructors reported at RMP for that establishment. Third, analyze the distribution of the descriptive factors (i.e.,quantitative score) of the top-30 universities (23 USA and 7 Canadian).

IV. THREATS TO THE VALIDITY AND FUTURE WORK

Construct Validity: The feature selection methodology cannot distinguish between explicit and implicit features that impact student learning goals. Implicit features include but are not limited to a sense of humor, amazing, hot, instructor charisma, and entertainment value. Whereas explicit features are clarity, helpfulness, homework, understanding, interesting material, knowledgeable, organized and attendance.

External Validity: One of the data sources used in the paper is RMP. The paper cannot definitively prove or disprove whether online student ratings are valid. At times, reviews can also take too much into account professor personalities over their ability to educate. They may be affected by emotion. In most cases, the only reason anyone leaves a review is because they had a less-than-pleasant experience. To mitigate the threat, the methodology uses the teaching evaluations obtained from the author's department.

***Conclusion Validity*:** While the paper present a methodology to infer features and distinguish students opinion, the methodology does not reveal the underlying reason behind forming such opinions. To mitigate the threat, future research using sentiment topic recognition (STR) will be conducted to determine the most representative topics discussed behind each sentiment.

Future Work: Generalizability of the methodology is one of the most important extensions of future work. Integration of a custom-built sentiment classifier and an automatic ontology building functionality potentially through a combination of ontology learning techniques will be sought.

References

[1]. George A. Miller. 1995. WordNet: a lexical database for English. Commun. ACM 38, 11 (November 1995), 39–41. DOI:https://doi.org/10.1145/219717.219748

[2]. M. Hu and B. Liu, "Mining opinion features in customer reviews," in AAAI, 2004, pp. 755-760.

[3]. N. Gupta, S. Chandra, Product Feature Discovery and Ranking for Sentiment Analysis from Online Reviews, Journal of Systems Science and Systems Engineering, pp 542-55, 2013.