FINE-GRAINED APPROACH OF SENTIMENT ANALYSIS FOR FACULTY PERFORMANCE EVALUATION

[1]Jomar C. Llevado and Jocelyn B. Barbosa[*,1]

[1] Department of Information Technology, College of Information Technology and Computing, University of Science and Technology of Southern Philippines, Republic of the Philippines
For Correspondence; Tel. +63 917 145 3957, Email: [*,1]jocelyn.barbosa@ustp.edu.ph

ABSTRACT: *Faculty performance evaluation is the conventional method used among Higher Education Institutions (HEI) to gain needed insights as to the basis for important decisions that may involve faculty tenure, promotion, and scholarships. Most often, these evaluations utilized numerical ratings to measure the faculty performance while textual feedback may have no bearing at all. As such, textual feedback has been widely adopted to measure faculty performance through text mining. However, as student's textual comments or feedback are subjective in nature, they may not capture all insights that are normally found in a standard survey questionnaire. Thus, faculty performance evaluation, which are solely based on text comments may not in turn give reliable results. Hence, the need to combine numerical ratings and textual feedback as a measure of the overall faculty performance. In this study, we introduce an innovative approach in combining numerical ratings and textual feedback in measuring the overall faculty performance. We leverage the use of Fine-Grained Sentiment Analysis to classify the textual feedback into five sentiment scores. We then assigned a numerical weight score to each polarity to calculate the overall faculty performance combined with the numerical ratings. We developed a web-based faculty evaluation system to streamline the collection of both the numerical and textual evaluation data using the instrument used by all State Universities and Colleges in the Philippines (i.e., QCE NBC 461) integrating the sentiment analysis module created. The system was deployed in University of Science and Technology of Southern Philippines. Experiments reveal that our proposed approach is efficient and provides effective results in terms of the system's usability and functionality.*

Keywords: Fined-Grained Sentiment Analysis, Faculty Performance Evaluation, Numerical Rating, Text Sentiment

1. INTRODUCTION

Faculty performance evaluation has become ubiquitous in most higher education institutions worldwide [1]. It has become one of the necessary parts of the education management process that provides additional information to academic administrators in making significant decisions related to faculty tenure, promotion, and scholarships [2-4]. Moreover, it can be a source of essential information towards improving pedagogical practices in most academic institutions [3].

Traditionally, these evaluations used questionnaire-based forms [1, 5, 6] administered at the end of a semester. The most commonly used form of student rating instrument combines Likert scale answers to statistically relevant questions with one or more open-ended questions [5]. Most often, numerical ratings or student ratings are used to assess a faculty member's teaching ability [7, 4]. In contrast, the textual feedback has little to no bearing in rating the overall faculty performance [7, 4]. On the other hand, textual feedback has the flexibility that questionnaire-based questions cannot address [5, 8]. As a result, textual data has been widely adopted in education through text mining [9-11]. According to Himelein, M. J [12], student textual feedback appears to equate with numerical scores, but outlier statements, on the other hand, are not rare and often contain strongly voiced opinions, possibly amplifying their effect. As students' textual comments or feedback are subjective, they may not capture all insights generally found in a standard survey questionnaire. Although qualitative data cannot thoroughly answer all evaluation questions, they can be combined with quantitative instruments to clarify overall faculty performance [13].

With Sentiment Analysis and Opinion Mining, the unstructured textual student evaluations can be analyzed and then extract the different polarities if it is positive, negative, or neutral [14-16]. Some current studies on faculty evaluation systems [6, 17-18] utilize Sentiment Analysis extracting Coarse-grained sentiment. These studies only provide a simple measure of the degree of polarization of opinion on a given subject or keyword. They do not provide in-depth analysis at the fine-grained level, which may reveal additional and precise details when applied to analyzing and processing the faculty performance evaluation given by students. Compared to coarse-grained analysis, in-depth analysis at the fine-grained level may provide specific details since emotions are articulated on a single or several topics within or through sentences [4, 19].

In this study, we introduce an innovative approach of combining numerical and textual feedback to measure the overall performance of the faculty. We leverage Fine-grained Sentiment Analysis using a Rule-based approach to extract five sentiment classifications: very positive, positive, neutral, negative, and very negative, from the textual feedback. We integrated all of the faculty performance evaluation process through a Web-Based Faculty Performance Evaluation System that allows extraction sentiments from unstructured text feedback of students. The objective is to develop a web-based faculty performance evaluation that would streamline the process of collecting the faculty performance evaluation data and combine both textual feedback and numerical ratings as basis for overall faculty performance. Specifically, our objectives include: (1) To design and develop a web-based faculty performance evaluation system; (2) To extract fine-grained sentiments from the textual feedback of students, and (3) To evaluate the efficacy of the web-based faculty performance evaluation system. The research is limited to the students' English verbatim as a baseline for further study. Additionally, this sentiment analysis excludes emojis, icons, symbols, and other features not explicitly specified in the objectives.

MATERIALS AND METHODS

A. System Architecture of the Web Based Faculty Performance Evaluation System

Our system's architecture represents the mechanism by which users communicate with our system. Figure 1 illustrates the design of our system architecture, which includes various web technologies. The data is accessible from the database server via computers and smartphones. Embedded in the system process is the automatic classification of verbatim comments. Additionally, we design an API in our system to facilitate the CRUD (Create Read Update Delete) operations required for data manipulation.

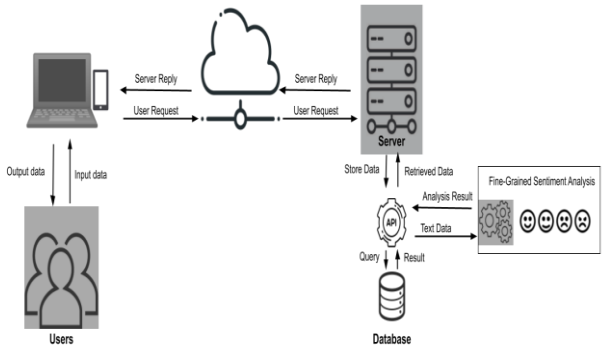

Figure 1. System Architecture of Web-Based Faculty Performance Evaluation System

B. Context Diagram of the Web Based Faculty Performance Evaluation System

The context diagram is relevant during the design and implementation of the system because our study includes data collection of assessment scores and verbatim. Our context diagram presents the overview of the web-based faculty performance evaluation system. Figure 2 denotes the interface between the system and its environment, indicating the entities with which it interacts. The diagram depicts that management or admin handles the updates of four entities (e.g., Faculty, Students, Evaluation Form, Course Subjects). The system can automatically generate reports (e.g., Faculty Evaluation report, Periodic Reports) that simplifies the manual compilation of student ratings and interpretation of the students verbatim. The system utilized QCE NBC 461 as the instrument used in evaluation by the students.

C. Fine-Grained Sentiment Analysis Process of Text Feedback

This part describes the technique used to apply fine-grained sentiment analysis on textual feedback of faculty performance evaluation. Figure 3 presents the systematic process of classifying the student textual feedback. We define text's polarity using fine-grained classification into five categories: very positive, positive, neutral, negative, or very negative using Rule-based

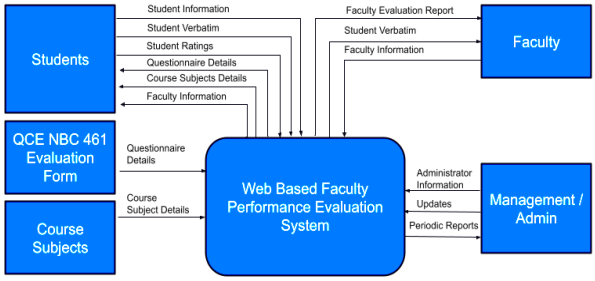

Figure 2. Context Diagram of Web-Based Faculty Performance Evaluation System

approach. Additionally, we further discuss the process of fine-grained sentiment analysis from text processing up to sentiment classification.

Text Preprocessing. The goal of this process is to eliminate the irrelevant and noisy text data. In this study, the preprocessing steps we apply are the following:

a. *Tokenization*. This process breaks down paragraphs, sentences, or phrases into fragments of words or phrases. For example, "He is the best instructor" will be converted into [He, is, the ,best, instructor] tokens.
b. *Case Conversion*. This process changes the text into lowercase or uppercase form of the text input.
c. *Special Characters Removal*. This process removes unwanted characters from the text input. Example of these are punctuation (e.g.,!;) and special characters (e.g., #$@_%*!) and etc. These characters are considered noise from the text data.
d. Stop word removal. This process removes words that has no significant meaning such as "is", "the","an", and etc.

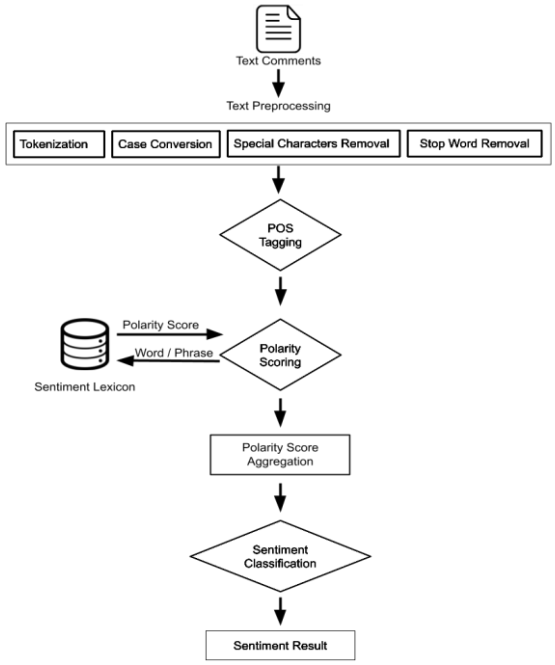

Figure 3. Fine-Grained Sentiment Analysis Process Flowchart

Part of Speech (POS) Tagging.
The process of classifying words into their parts of speech and labeling them accordingly such as verb, adjective, etc. (see Table 1). For example, the sentence "He is the best instructor". Using abbreviation tag like NN for noun, VB for verb, JJ for adjective, PP for preposition and other part of speech as tags. We can tag each of the pre-processed text of its corresponding POS tag which may play significant part during the polarity scoring as their word valence may differ base on the type of POS in the Sentiment Lexicon.

Table 1. Example of POS Tagging a word

Word	POS Tag
He	NN
is	VB
the	DET
best	JJ
instructor	NN

Sentiment Lexicon. A sentiment lexicon is a list of terms (alternatively called polar or opinion words) classified according to their sentiment orientation, that is, whether they are positive or negative. Given the time and resources it will take to generate our own annotated valence-based sentiment lexicon we leverage the use of VADER Sentiment Lexicon to identify a much fine-grained sentiments using valence-based lexicons.
Polarity Tagging. Polarity tagging is the process that analyzes the text data and tags the word with the valence score as positive, negative, neutral from the sentiment dictionary. *Valence score or Polarity score* is a score assigned to the word base on its intensity of positive or negative emotion. Using the available sentiment lexicon, we extract the valence score of each word from the text data.
Polarity Score Aggregation. In this step, the polarity or valence score of the tagged words is aggregated base on their prior orientation which is either positive, negative or neutral. We then aggregated the overall valence score of the text data by adding all the result from the average of prior orientation valence. In this study, we use a metric in between of -1 and +1 to measure the sentiment score. We then use a formula to normalize the overall valence score between -1(very negative) and +1 (very positive). We applied the following formulas to calculate the average valence scores of the prior orientation, the overall valence score, and the normalization of the overall valence score.

a. To calculate the average polarity scores of the prior orientation of the words in the text data. We apply the following formula in equation (1), where is the mean of all polarity valence score; x is the summation of the polarity valence score; and n is the total number of words with the same polarity orientation in the text data.

$$\overline{p} = \frac{\sum(x)}{n} \qquad (1)$$

b. To calculate the overall valence score. We apply the following formula in equation (2) to calculate the overall polarity score; *pos* is the mean of all word positive polarity score; *neg* is the mean of all negative polarity score; *neu* is the mean of all neutral polarity score; and *n* is the total number of tagged words in the text data.

c. To normalize the compound valence score. We apply the following formula in equation (3), where x is the overall valence score and alpha is set to be 15 which approximates the maximum expected value of x. This formula yields a value in between -1 and +1.

$$\text{Sentiment Score} = \frac{x}{\sqrt{x^2 + \alpha}} \qquad (3)$$

Sentiment Classification. To classify the polarity of the text evaluation data we use the normalized aggregated valence scores and use a metric to classify our text feedback into very positive, positive, neutral, negative, and very negative.

3. RESULTS AND DISCUSSION

In this section, we present some sample screen shots of our web-based faculty performance evaluation system, which includes modules for the student evaluation, dashboard for admin, dashboard for faculty, and other important features.
With this system, the tedious work involves processing the data collection of student ratings and verbatim, analysis of the result, and compilation of periodic reports. Figure 4 shows the sample screen shot of the landing page of the faculty evaluation system with the login interface where users must enter a valid username and password. We design our system with several access levels from student being the lowest level to the system admin being the highest. Depending on the access levels, users can perform create, search and update tasks. Figure 5 presents the admin dashboard of the system, where system settings configuration and other admin task can be executed. While Figure 6 shows the student dashboard where students can access the evaluation for their subjects and the evaluation form used in the evaluation.

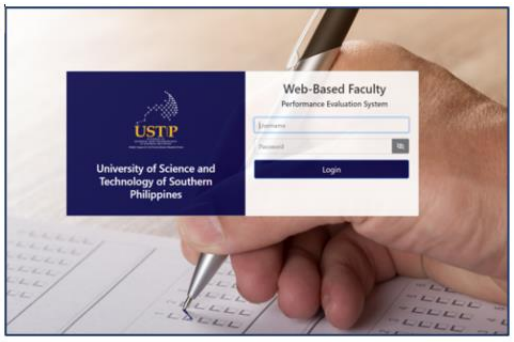

Figure 4. Sample screen shot of the landing page of the system

Figure 5. Sample screen of the admin dashboard

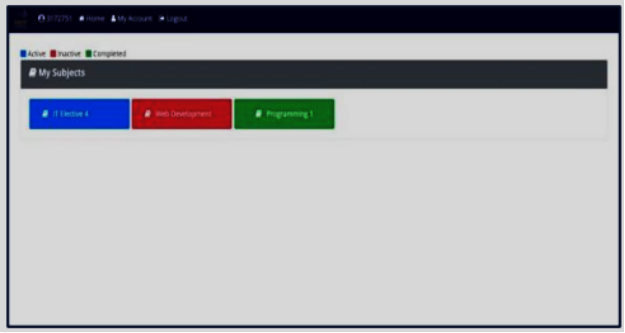

Figure 6. Screen hot of the student dashboard and the form used in the evaluation.

3.1 Student Verbatim Classification

To test the accuracy of our classifier, we extracted our test data from the first semester of the current school year. We curated and manually labeled the test data of its polarity as either positive and negative—the test data comprised 263 labeled as positive comments and 121 labeled as negative. The low sample set of test data resulted primarily from redundant student comments and mixed English and Bisaya word, so we trimmed and cleaned our test data with a sample set used in our classification.

Table 2 shows the sample result from the classified student verbatim. It shows the sentiment score for each verbatim and the classification of its polarity. We use 0.5 or higher as our threshold for the very positive polarity, -0.5 or higher for our very negative polarity, 0 for neutral, 0.5 lower and higher than 0 for the positive polarity, and lower than 0 and higher than -0.5 for negative; this was based on VADER recommended threshold [19].

Table 2. Example of classified fine-grained polarity of the student verbatim

Student Verbatim	Sentiment Score	Polarity Classification
Thank you so much ma'am for being one of the best teachers to us.	0.7717	Very Positive
Ma'am is teaching based on his experiences	0	Neutral
Thank you for the first semester learning experience.	0.36	Positive
Does not give considerations even with proper reasons and uses "mean" and "insulting" words towards students.	-0.49	Negative
She insulted me when I was asking questions about our class and also embarrassed my classmates during out class lessons.	-0.79	Very Negative

Our test data was comprised of 68% labeled as positive verbatim and 32% labeled as negative verbatim. With the 68% test data, 65.2% of the positive student verbatim was correctly identified as positive by our classifier; further fine-grained classification shows 41.3% was classified as very positive and 23.9% as positive. From the 32% labeled as negative student verbatim, 16.4% was identified as negative by our classifier; further fine-grained classification shows that 12.47% was classified negative and 3.9% was classified very negative sentiment. With our classifier, we yield an 81.6% accuracy based on the experiments we conducted. Figure 7 shows the distribution of the polarity from the test data we classified.

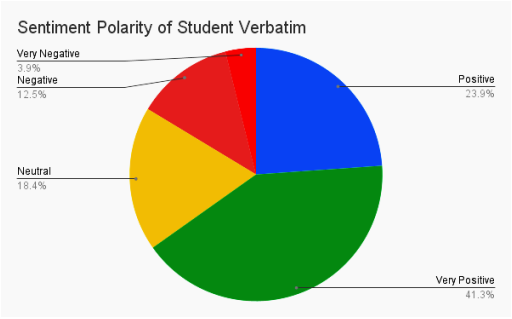

Figure 7. Distribution of fine-grained sentiment polarity of student verbatim

In what follows, we also discussed the performance of our proposed approach addressing a classification problem on the collected students' verbatim or textual comments per subject per faculty. Furthermore, the survey using S.U.S instruments was conducted to evaluate the system's usability.

3.2 System Usability Evaluation

According to ISO standard ISO 9241, the usability of a system can only be determined by understanding the context in which it is used (i.e., who uses it, what they use it for, and the setting or workspace in which they use it).

Additionally, usability metrics include the following: 1. Effectiveness (can users accomplish their goals successfully); 2. Efficiency (how much time and resource is spent to accomplish those goals); and 3. Satisfaction (was the experience satisfactory). To assess our system's usability, we use the System Usability Scale (SUS) [20], a systematic metric for evaluating the usability of a web-based system or other software application. SUS is a ten-item attitude Likert scale used in systems engineering that provides a consolidated perception of subjective usability tests. SUS is an extremely useful quantitative instrument for those attempting to improve the user experience. SUS uses a short, 10-item questionnaire administered at the end of a usability test to calculate usability score of the system. Users respond to each question using a 5-point scale from "Strongly disagree" to "Strongly Agree" [21]. We administered survey questionnaires based on the SUS's ten-item questionnaire. One hundred fifty-eight (158) respondents were asked to score each question on a scale of 1 to 5, with 1 indicating the least (Strongly Disagree) and 5 indicating the highest (Strongly Agree) grade. The survey results indicate that correspondents are satisfied with the system and are more likely to recommend it. Figure 6 presents the findings of a usability evaluation conducted using the SUS instrument on the device. The survey yields 75.6% aggregated SUS score, which implies a high acceptability rating from the users.

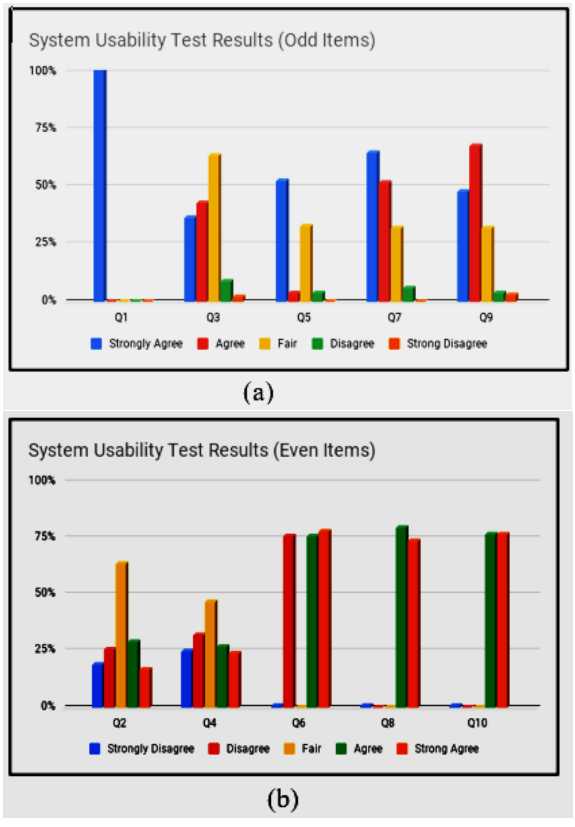

Figure 6. SUS survey results. (a) Odd items results; and (b) Even items results.

4. CONCLUSION AND RECOMMENDATIONS

The researchers concluded that the development of Web-Based Faculty Performance Evaluation greatly enhances the process of timely and valuable faculty evaluations. With our novel approach in combining both numeric rating and textual feedback, we added a new value to the overall faculty performance and put a premium on student textual feedback as part of the faculty evaluation process. Based on the findings, the following recommendations may be considered for future research work: (1) Fine-Grained Sentiment Analysis of mixed language in the text data; (2) More datasets to use for modelling and test data; and (3) Inclusion of emojis and symbols on a Fine-Grained Sentiment Analysis.

REFERENCES

[1] Darwin, S. (2016a). The Emergence of Student Evaluation in Higher Education. Student Evaluation in Higher Education, 1–11. https://doi.org/10.1007/978-3-319-41893-3_1

[2] Alkathiri, M. S. (2021). Decision-making by Heads of Academic Department using Student Evaluation of Instruction (SEI). International Journal of Learning, Teaching and Educational Research, 20(2),

[3] Benton, S., W. Cashin, and K. Manhattan. 2012. Student ratings of teaching: A summary of research and literature. IDEA Center. http://www.ideaedu.org

[4] Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. Studies in Educational Evaluation, 54, 94–106. https://doi.org/10.1016/j.stueduc.2016.12.004

[5] Kumar, A., & Jain, R. (2018). Faculty Evaluation System. Procedia Computer Science, 125, 533–541. https://doi.org/10.1016/j.procs.2017.12.069

[6] Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-Based Sentiment Analysis of Teachers' Evaluation. Applied Computational Intelligence and Soft Computing, 2016, 1–12. https://doi.org/10.1155/2016/2385429

[7] Benton, S. L., & Ryalls, K. R. (2016). Challenging misconceptions about student ratings of instruction (IDEA Paper)

[8] Wongsurawat, W. (2011). What's a comment worth? How to better understand student evaluations of teaching. Quality Assurance in Education, 19(1), 67–83. https://doi.org/10.1108/09684881111107762

[9] Cook, J., Chen, C., & Griffin, A. (2019). Using Text Mining and Data Mining Techniques for Applied Learning Assessment. Journal of Effective Teaching in Higher Education, 2(1), 60–79. https://doi.org/10.36021/jethe.v2i1.39

[10] Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(6), 1. https://doi.org/10.1002/widm.1332

[11] Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Hammout, N., Nakamura, E. L., Escamilla, J., & Hosseini, S. (2020). Impact of students evaluation of

teaching: a text analysis of the teachers qualities by gender. International Journal of Educational Technology in Higher Education, 17(1), 49. https://doi.org/10.1186/s41239-020-00224-z

[12] Himelein, M. J. (2018, November 29). Pitfalls of Using Student Comments in the Evaluation of Faculty. Academic Briefing | Higher Ed Administrative Leadership. https://www.academicbriefing.com/human-resources/faculty-evaluation/pitfalls-of-using-student-comments-evaluation-of-faculty/

[13] Harper, S. R., & Kuh, G. D. (2007). Myths and misconceptions about using qualitative methods in assessment. New Directions for Institutional Research, 2007(136), 5–14. https://doi.org/10.1002/ir.227

[14] Liu, B. (2012). Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[15] Liu, B. (2020). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions (2nd ed.). Cambridge University Press. https://doi.org/10.1162/COLI_r_00259

[16] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

[17] Balahadia, F. F., & Comendador, B. E. V. (2016). Adoption of Opinion Mining in the Faculty Performance Evaluation System by the Students Using Naïve Bayes Algorithm. International Journal of Computer Theory and Engineering, 8(3), 255–259. https://doi.org/10.7763/ijcte.2016.v8.1054

[18] Balahadia, F. F., Fernando, M. C. G., & Juanatas, I. C. (2016). Teacher's performance evaluation tool using opinion mining with sentiment analysis. 2016 IEEE Region 10 Symposium (TENSYMP), 1. https://doi.org/10.1109/tenconspring.2016.7519384

[19] Wang, Z., Chong, C. S., Lan, L., Yang, Y., Beng Ho, S., & Tong, J. C. (2016). Fine-grained sentiment analysis of social media with emotion sensing. 2016 Future Technologies Conference (FTC), 1. https://doi.org/10.1109/ftc.2016.7821783

[20] Brooke J. SUS: A Retrospective. Journal of Usability Studies. 2013; 8(2):29–40