Unraveling Views of Students Towards Computer Programming A Sentiment Analysis and Latent Semantic Indexing Application

Niel Francis Casillano

Abstract*: Programming is a fundamental skill a computing student must master. It requires an excellent and correct understanding of logical and abstract concepts. Due to this, students find a hard time dissecting and understanding programming problems. This paper focused on unraveling the views and sentiments of students towards computer programming. The researcher utilized a machine learning tool to analyse and visualize the corpus of documents containing the views and sentiments of students. VADER model was used to analyze the sentiments of the students and Latent semantic indexing topic modeling was utilized to generate topics from the corpus of documents. It was determined that most students have a negative sentiment towards programming subjects. The topic modeling showed that the underlying themes were generally talking about the difficulties and challenges students are experiencing when dealing with programming subjects. It was also observed that some students are using coping mechanisms and finding new learning methodologies to solve programming tasks given to them. The result of this research can be utilized as inputs in the development of a teaching model for programming.

***Index Terms*: LSI, Latent Themes, Topic Modelling, Programming, Sentiment Analysis**

I. INTRODUCTION

The technology revolution and the era of the knowledge-based economy have increased the demand of people who are graduates of computer related programs such as computer science, engineering, and information technology [1]. It has been established through research that there is a high demand for software engineers and programmers in the industry and that programmers are critical requirements in the success of a company [2]. It is essential for Information Technology educators to develop effective learning strategies and methodologies that will help prepare students who are pursuing a career in computer programming [1]. Programming has been deemed as an important instrument in improving higher-order thinking skills of a student [3]. It is considered as a potent tool with which student can explore and solve computing problems by editing, analyzing, evaluating and explaining their chain of thoughts clearly [4]. However, programming is a difficult subject to learn. Programming has been considered as a complex subject matter and has been categorized as one of the seven grand challenges in computing [5]. Entry level programming students are commonly encountering problems in reading, tracking, writing and designing simple programs. This makes students depend on the internet and their peers to complete a programming problem and they consequently become lazy and less courageous to learn and expect mediocre grades from their professors [6]. The inability to immediately grasp programming concepts poses a challenge for professors to utilize the best and the most appropriate teaching strategy that will provide students with the most effective learning environment [7].

Eastern Samar State University, the sole state university in the province of Eastern Samar, Philippines, is offering Computer Science and Information Technology programs. The researcher being a programming lecturer has observed a very high failure rate in programming subjects. Students tend to submit an unfinished version of their programming tasks and usually rely on their peers to complete their problems. It has also been observed that students fail to start coding without the assistance notes or a pre-made program. Therefore, it is imperative to investigate and identify the root cause of the problem in order to properly address it and aid the students in improving their programming skills. In this paper, the researcher will extract and determine the views and sentiments of students towards computer programming using Valence Aware Dictionary for sEntiment Reasoning (VADER) model and a topic modeling technique called Latent Semantic Indexing.

II. OBJECTIVES

This study focused on exploring documents that were collated from Information Technology students. The documents contain the sentiments and views of students towards computer programming. The collection of data served as an input to a machine learning tool to produce sentiment analysis and topic modeling results.

Specifically, this study aimed to achieve the following:

1. Determine the frequently occurring words in the collection of documents through a word cloud.
2. Identify the Over-all sentiment of the collection of documents using the Valence Aware Dictionary for sEntiment Reasoning (VADER) model and a Heat Map
3. What are the hidden issues or topics prevalent in the different articles as produced by employing Latent Semantic Indexing (LSI).

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author

Niel Francis B. Casillano*, Department of Information Technology, Eastern Samar State University, Philippines

III. LATENT SEMANTIC INDEXING

Latent Semantic Indexing (LSI) is a technique which aims to overcome some of the limitations that commonly occurs in a conventional Vector Space Model (VSM) [8][9]. VSM uses the process of collecting word in groups called bag-of-words representation of documents, the set of documents is represented by a matrix A = [aij] ∈ Rt×d, where aij represents the number of times the terms I appear in the document j. The variable d represents the number of documents in the set of documents while t represents the number of terms. Using the formula, the document becomes a column vector and the user's query is presented as a vector of the same dimension. The similarity of the two vectors is measured as the cosine of the angle between the two vectors. An output containing the list of documents ranked in decreasing order of similarity is returned for every query [10]. LSI on the other hand, Is a variation of VSM that looks into the dependencies of words by assuming that there are some hidden or "latent" structure in usage of words within the entire set of documents [8]. LSI is used for dimensionality reduction through transforming the original terms documents vector space into a new system of conceptual topics. Instead of tagging the documents as vectors of independent words, the documents and terms are projected onto a low-dimensional space of concepts. LSI utilizes a shortened Singular Value Decomposition (SVD) applied to the aforementioned matrix A [10]. The SVD of A is denoted as $A = USV^T$ where U is a t × m orthonormal matrix ($U^TU = I_m$) whose columns define the left singular vectors, V is a d × m orthonormal matrix ($V^{\ T} V = I_m$) whose columns define the right singular vectors, and S is a m × m diagonal matrix containing the singular values of A decreasingly ordered along its diagonal: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_m = 0$, where r = rank(A). This decomposition is unique up to making the same permutations of columns of U, elements of S and columns of V (rows of V^T)[10].

IV. SENTIMENT ANALYSIS

VADER (Valence Aware Dictionary for sEntiment Reasoning) is a sentiment analysis model used to determine the polarity (positive/negative) and intensity (strength) of the emotion of a corpus or set of documents. VADER, introduced in 2014, uses a human-centric approach that combines qualitative analysis and empirical validation by using human evaluators and the crowdsourcing. To quantify the emotion of a word, VADER relies on a dictionary which maps words to emotion intensities called sentiment scores. The score of a text is computed by summing up the intensity value of each word in a corpus of documents. The scale used in measuring the emotion intensity ranges from -4 to +4, where -4 is the largest negative score (most negative) and +4 as the most positive. Meanwhile, 0 (midpoint) is considered neutral. A normalization formula is applied to the overall emotion intensity to map the score within a range of -1 to +1 [11][12].

V. RESEARCH METHOD

The research commenced with the importing of documents using an open source machine learning tool called Orange [13] as shown in figure 1. The documents were gathered from All BS Information Technology Third Year Students (n=21) of Eastern Samar State University – College of Computer Studies. The respondents were chosen because of their intense exposure to programming subjects versus the lower year levels. The students were asked the question *"What are your views and sentiments towards computer programming as a course?"* The documents were then converted into a digital format and were collated as one dataset. The collection of documents underwent pre-processing to remove all unnecessary words, punctuations, and elements. The pre-processed documents were fed to a bag of words model to create a corpus with word counts for each data instance and a sentiment analysis model called VADER. The generated corpus served as input to generate a word cloud that visualizes the most frequently occurring words. The same corpus was also used to generate a set topic by employing Latent Semantic Indexing Model.

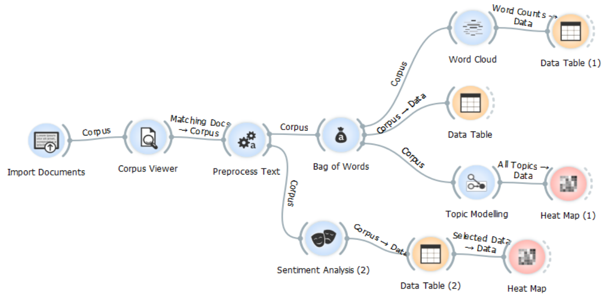


Figure 1. Research Framework

VI. RESULTS AND DISCUSSION

The word cloud that was generated from the corpus of documents showed the most frequently used terms by the students in their respective answers (as shown in figure 2). It was observed that the most predominant words in the corpus of documents were the words *programming*, *subject*, *difficult*, *hard*, *learn*, *understand*, and *challenging*. This probably implies that the students find a hard time understanding programming subjects and consider programming as a challenging field.

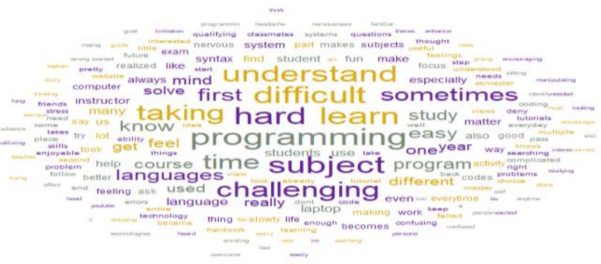
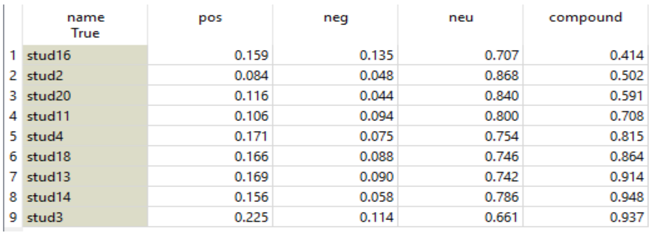


Figure 2. Word Cloud

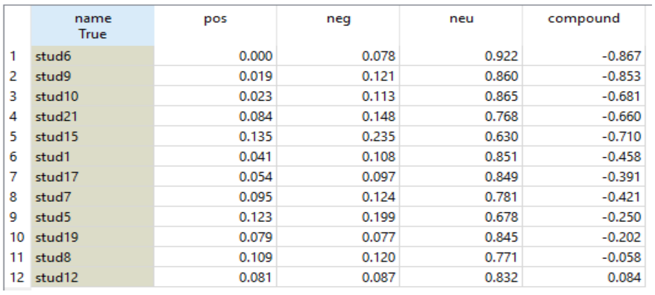
To analyze the sentiment of each document, the corpus was fed to a Sentiment analysis widget that utilizes the VADER model. It can be observed in Table 1 that documents *stud16*, *stud2*, *stud20*, *stud11*, *stud4*, *stud18*, *stud13*, *stud14*, and *stud3* was evaluated to be positive documents. This would infer that students who wrote the aforementioned documents felt a positive sentiment towards their experience on computer programming subjects. Meanwhile, documents *stud6*, *stud9*, *stud10*, *stud21*, *stud15*, *stud1*, *stud17*, *stud7*, *stud5*, *stud19*, *stud8*, *stud12*, shown in table 2, were tagged as negative documents based on the compound value of its sentiments. This would imply that the students used several negative words in their respective documents which can be attributed to the difficulty they experience in taking programming subjects. This data is in consonance with the result of the word cloud shown in figure 2.

Table 1. Documents with positive sentiment analysis results



	name True	pos	neg	neu	compound
1	stud16	0.159	0.135	0.707	0.414
2	stud2	0.084	0.048	0.868	0.502
3	stud20	0.116	0.044	0.840	0.591
4	stud11	0.106	0.094	0.800	0.708
5	stud4	0.171	0.075	0.754	0.815
6	stud18	0.166	0.088	0.746	0.864
7	stud13	0.169	0.090	0.742	0.914
8	stud14	0.156	0.058	0.786	0.948
9	stud3	0.225	0.114	0.661	0.937

Table 2. Documents with negative sentiment analysis results



	name True	pos	neg	neu	compound
1	stud6	0.000	0.078	0.922	-0.867
2	stud9	0.019	0.121	0.860	-0.853
3	stud10	0.023	0.113	0.865	-0.681
4	stud21	0.084	0.148	0.768	-0.660
5	stud15	0.135	0.235	0.630	-0.710
6	stud1	0.041	0.108	0.851	-0.458
7	stud17	0.054	0.097	0.849	-0.391
8	stud7	0.095	0.124	0.781	-0.421
9	stud5	0.123	0.199	0.678	-0.250
10	stud19	0.079	0.077	0.845	-0.202
11	stud8	0.109	0.120	0.771	-0.058
12	stud12	0.081	0.087	0.832	0.084

To visualize the result of the sentiment analysis, the data that was generated by the VADER model was fed to a heat map widget. The heat map visualizes the sentiments of each document through a set of colors. The color blue pertains to negative sentiments while yellow for positive. It can be noted in figure 5 that most of the documents are leaning towards the color blue which would infer that most of the student included in their answers words which were tagged as negative making a compound value of most documents negative.

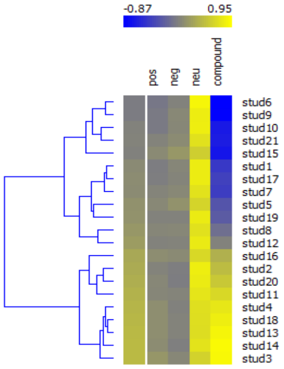
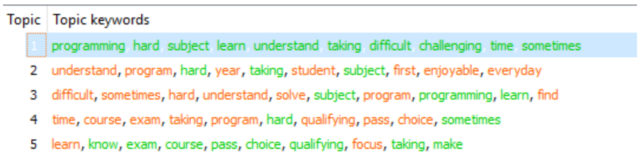


Figure 5. Heat Map for the Sentiment Analysis

To identify the latent topics within the corpus of documents, it was fed to the topic modeling widget that utilized latent semantic indexing. The model as shown in figure 6, was set to create five (5) topics which contained ten (10) words per topic. Based on the philosophical perspective of the researcher, latent themes for the generated topics were crafted (shown in Table 3). Topic 1 generally talks about the difficulty and challenges students are experiencing when taking up programming subjects. As a programming professor, it can be observed that students who have limited to no background in the realm of computing have very poor retention and understanding of programming concepts. Topic 2 talks about the students coping mechanism when dealing with programming subjects. Students who are productive and motivated to do programming problems treat programming as an enjoyable and rewarding subject. They consider programming as a challenge and tries different methodologies to solve the tasks given to them. Topic 3 talks about the student's difficulty in solving programming problems. This topic is highly related to topic 1. A student who has a limited understanding of programming concepts is generally unable to finish the programming tasks. This can be attributed to their inability to dissect the problem which leads to their inability to produce a solution for the problem. Topic 4 talks about programming problems related to exams. A student who is unable to finish simple programming activities becomes anxious and less motivated when dealing with major exams. Students tend to panic which results to them submitting an unfinished version of their programming tasks. Topic 5 talks about the students drive and motivation to pass programming subjects. While other students consider programming as a difficult subject, others are finding ways to gradually learn and eventually pass the subject. Some student who is not familiar with programming concepts utilize the trial and error method to successfully a find a solution for a given problem. Others try to understand existing codes and apply it to the problems they are working on.



Topic	Topic keywords
1	programming, hard, subject, learn, understand, taking, difficult, challenging, time, sometimes
2	understand, program, hard, year, taking, student, subject, first, enjoyable, everyday
3	difficult, sometimes, hard, understand, solve, subject, program, programming, learn, find
4	time, course, exam, taking, program, hard, qualifying, pass, choice, sometimes
5	learn, know, exam, course, pass, choice, qualifying, focus, taking, make

Figure 6. Topics Generated using LSI

Table 3. Latent Themes of the Topics generated through LSI

Topic	Latent Theme
1	Student's difficulty in understanding programming concepts
2	Student's coping mechanisms in dealing with programming problems
3	Student's difficulty in solving programming problems
4	Programming Problems relating to Examinations in Programming
5	Student's drive and motivation to pass programming subjects

VII. CONCLUSION

Programming has been dubbed as one of the most difficult courses in the field of computing [5]. To identify the views and sentiments of students towards programming, a corpus of documents were collated from the students. The documents were fed to a sentiment analysis and topic modeling module. It was determined that most students have a negative sentiment towards programming subjects. The topic modeling showed that the underlying themes were generally talking about the difficulties and challenges students are experiencing when dealing with programming subjects. It was also observed that some students are using coping mechanisms and finding new learning methodologies to solve programming tasks given to them.

REFERENCES

1. Özmen, B., & Altun, A. (2014). Undergraduate Students' Experiences in Programming: Difficulties and Obstacles. Turkish Online Journal of Qualitative Inquiry.
2. Ahmad Rizal Madar, Nurliana Musa, and YahyaBuntat, ―Kesan model pembelajaranberasaskankaedahpenyelesaianmasalahkeataspelajarberbezagayakognitifdankemahiranlogik, (The effect of using learning model based on problem solving method on students with different cognitive style and logic ability) in 2nd International Malaysian Educational Technology Convention, 2007.
3. Papert, S. (1991). Mindstorms: Children, computers and powerful ideas. Athens: Odysseas Publications (in Greek). C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.
4. diSessa, A.A., & Abelson, H.(1986). Boxer: A reconstructible computational medium. Communications of the ACM, 29(9), 859–868.
5. McGetrick, R. Borle, R. Ibbett, J. Llyod, G. Lovegrove, and K. Mander, ―Grand challenges in computing: Education-A Summary, The Computer Journal, vol. 48, no. 1, pp. 42–48, Jan. 2005.
6. Aziz Deraman, ― Teaching software factory: An approach to increasing programming skills) in Proceeding of Science Programming Workshop: Teaching and Learning in Malaysia (ATUR'03), 2003, pp. 1–7.
7. S. Mohorovicic and V. Strcic, "An Overview of Computer Programming Teaching Methods," pp. 1 – 6, (n.d).
8. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R. (1990): Indexing by latent semantic analysis. — J. Amer. Soc. Inf. Sci., Vol. 41, No. 6, pp. 391–407.
9. Salton G. (1971): The SMART Retrieval System: Experiments in Automatic Document Processing. — Englewood Cliffs, NJ: Prentice Hall.
10. Moldovan, A., Bot, R., & Wanka, G. (2005). Latent Semantic Indexing For Patent Documents. Int. J. Appl. Math. Comput. Sci.,, 551-560
11. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Association for the Advancement of Artificial Intelligence.
12. Calderon, P. (2017, April 10). http://datameetsmedia.com/vader-sentiment-analysis-explained/. Retrieved from http://datameetsmedia.com.
13. Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research 14(Aug):2349−2353.

AUTHORS PROFILE

Niel Francis B. Casillano obtained his BSc in Information Technology at Leyte Normal University, Philippines. He also holds a Master of Arts in Teaching Computer Science. He is currently working as a Fulltime Instructor and a Research and Extension Coordinator at Eastern Samar State University. His research interests are the field of software engineering, data mining, data security, e-learning and software evaluation.