Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol.8, No.1, 2018
www.iiste.org

Mining Students' Messages to Discover Problems Associated with Academic Learning

Mensah Kwabena Patrick Akobre Stephen
Department of Computer Science, University for Development Studies, P. O. Box 24, Navrongo, Ghana

Abstract
WhatsApp has become the preferred choice of students for sending messages in developing countries. Due to its privacy and the ability to create groups, students are able to express their "feelings" to peers without fear. To obtain immediate feedback on problems hindering effective learning, supervised learning algorithms were applied to mine the sentiments in WhatsApp group messages of University students. An ensemble classifier made up of Naïve Bayes, Support Vector Machines, and Decision Trees outperformed the individual classifiers in predicting the mood of students with an accuracy of 0.76, 0.92 recall, 0.72 precision and 0.80 F-score. These results show that we can predict the mood and emotions of students towards academic learning from their private messages. The method is therefore proposed as one of the effective ways by which educational authorities can cost effectively monitor issues hindering students' academic learning and by extension their academic progress.
Keywords: WhatsApp; Sentiments; Ensemble; Classification; Naïve Bayes; Support Vector Machines.

1. Introduction

Much of the available research on sentiment analysis (Ortigosa, Martín, & Carro, 2013), (Korenek & Marián, 2013), (Chikersal, Poria, Cambria, Gelbukh, & Siong, 2015), (Carchiolo, A. L. B, & Malgeri, 2015), (Janssen, Tummel, Jeschke, & Richert, 2015) used data from major social media sites such as Twitter, Facebook and web-content such as blogs. Little or no attention has been given to messaging platforms such as WhatsApp which allows users to form groups. Messages to these groups can be sent and read only by members of the group, making it an appropriate medium for students to express their true emotions concerning their academic work such as learning. University students in Ghana create WhatsApp groups as suitable alternatives to mainstream social media for sharing their problems, experiences, and emotions. Students are able to express themselves freely because they believe the digital footprints they leave behind cannot be accessed by anyone outside of their group without permission. These messages are an enormous source of data from which we can gain insight into the learning process. Most of the educational data mining research focus on either the use of online learning data, course management systems data, or classroom technology data. A few concentrate on the use of Social Media Mining (Chen, Member, Vorvoreanu, & Madhavan, 2014) to understand student's sentiments with some using unstructured mobile data (messages) from student WhatsApp groups. The problem this paper attempts to solve can be stated as follows: Students encounter different type of problems which they find difficult to discuss with school authorities but less difficult to discuss informally with their colleagues. How do we get an idea about the core problems facing them in their studies? What type of mood do they express towards learning in the academic setting? Towards which aspects of the learning process do they express positive or negative emotions? Answers to these problems will be beneficial in formulating solutions for timely interventions. This paper aims at identifying students' sentiments using a cost-effective approach to serve as a basis to determine the mood of students towards every aspect of academic learning. It will also be fundamental to know if these emotions affect their academic performance and to enable students get the best out of their lecturers (Fig. 2).

About 70% of students in Ghana own a smart phone with WhatsApp installed and in use. They spend countless hours on their phones sending and receiving WhatsApp messages, making it one of the most used messaging platform in many developing countries (Yeboah & Ewur, 2014), (Khanna & Sambandam, 2015), (Mensa-Bonsu, 2016), (Dorwal, Sachdev, Gautam, & Jain, 2016). Cost and internet connectivity problems are lessened by the small amount of data required to run WhatsApp as opposed to Twitter and Facebook. Therefore, majority of the students prefer WhatsApp groups to mainstream social media.

WhatsApp messages were collected from four program options in one of the Universities in Ghana: BSc. Computer Science, BSc. Information Technology, BSc. Computing with Accounting (Level 300 and Level 400) and Diploma in Computer Science during the 2015/2016 academic year. Table 1 shows the total number of messages retrieved from each program option. Three binary classifiers and an Ensemble classifier based on majority voting were trained and used for the classification of the messages. Comparatively, the Ensemble classifier outperformed the individual classifiers and was used for further analysis. Fig. 1 depicts the workflow adopted in this paper.

The contributions of this paper can be summarized as follows:

- A demonstration that students' problems and emotions can be understood by mining cheaply sourced private messages to enable students get the best out of their lecturers and the system (Fig. 2.)
- It proposes a framework to mine student's attitudes to learning and during learning.

- To serve as a tool to timely and cost effectively identify troubled classes.

For purposes of this work, authors of messages were identified using the student's mobile number and identity number available in the student's management information system shown in Fig. 1.

Research shows that students prefer WhatsApp to alternate platforms despite its numerous negative effects on academic performance (Yeboah & Ewur, 2014). There is therefore the potential to discover the mood of students towards learning by mining these cheaply sourced data.

1.1 Educational Data Mining

Tools aimed at either improving teaching and learning (Westerfield, Mitrovic, & Billinghurst, 2015) or assessing how effective teaching and learning is (Zorrilla, Álvarez, & García-Saiz, 2015) are well established and widespread. However, they are costly, time consuming and untimely. Data mining is therefore, fast becoming a better alternative in terms of time and cost to surveys in education.

Educational data mining can be categorized into three broad application areas. The first category uses data mining to discover students who are at risk in order to prevent student attrition and improve retention. Chat mining has been shown to be one of the effective ways to improve educational learning environment (Wai, Chu, Wai, & Lam, 2015). Category two is an application of data mining to provide personal learning recommender systems. These systems can adapt to each student's learning requirements. Student's academic success and performance can be predicted as a result (Gowda, Baker, Corbett, & Rossi, 2013), (Arroyo, et al., 2014). Intelligent Multimedia adaptive systems in specific subjects have been proposed with a tutor that can adapt to meet the special needs of the learner (Walker, Rummel, & Koedinger, 2014). The third category in this area involves data mining to improve the usefulness of Course Management Systems. Machine learning algorithms have been applied on data from course management systems (CMS) to obtain excellent results. Algorithms have been applied to log data of web-based learning systems to predict students at risk in the early stages (Arroyo, et al., 2014) and to also find interesting patterns of student's online behavior. Online learning tools have been developed with the ability to mine student's difficulty during learning (Merceron & Yacef, 2003). Others are purposely designed to collect data on student learning experience for purposes of extracting usage patterns and student behavior during learning (San Pedro, d Baker, & Rodrigo , 2014), (Sabourin, Shores, Mott, & Lester, 2013). This can then be used to predict the performance of those students involved.

1.2 Sentiment Analysis in Education

A sentence may be a fact, an opinion, both, or even falsehood. Opinionated sentences can convey either positive or negative sentiments. Opinions people carry in many instances will influence their attitude (Mensah, 2015). Therefore, discovering the opinions of students in a program can be a pre-requisite for monitoring their progress in the program. Mining sentiments will indirectly discover attitudes including emotions of the student towards the course, its lecturer, the environment and colleagues. Sentiment analysis and opinion mining has been used to determine what people's opinion (Younis, 2015), (Janssen, Tummel, Jeschke, & Richert, 2015) are. The method can serve as feedback from students based on which authorities can act. Sentiment analysis has successfully been applied on ratings and textual responses of student evaluations of teaching, analysis of student questions and answer board for specific course like Computer Science, Massive Open Online Course (MOOC) Discussion Forums, teacher evaluation policies, and for evaluating Universities (Janssen, Tummel, Jeschke, & Richert, 2015). It is known that factors such as accommodation, medium of instruction, parent's educational background, family financial status, place of origin, personal habits, etc. can be detrimental to affective learning. The challenge then is how to cost-effectively determine in timely fashion which factor is specific to which student or group of students.

2. Materials and Methods

The mobile messaging application called WhatsApp is currently becoming the preferred choice of communication (Guler, 2016), [23], [24], (Seufert, Hoßfeld, Schwin, Burger, & Tran-gia, 2016)for most institutions compared to Twitter and Facebook due to cost related to data and simplicity of use. The application allows users to send secure private messages to their contacts on WhatsApp at a cheaper cost compared to other messaging platforms. To enable many people to have access to a common set of messages on the platform, a group of contacts can be created to share messages. It takes the initiative of one person of the group to create the group and add contacts as members. The group creator can act as the administrator or assign the task to all members of the group. Currently in Ghana, WhatsApp groups exist for teacher unions, student groups, market women, and can be found everywhere people come together to form a group. The main objective of these groups is to share information concerning the common agenda that brings them together.

The "Email chat" facility in the application was used to collect the chats from September 2015 to June 2016 which is one complete academic year. Messages not related to academic work were considered irrelevant and removed. Table 1 shows the number of messages collected from each of the groups. Messages collected were from higher level classes because they were at that moment doing their core departmental courses actually related to

their options. Lower level Students have always complained about borrowed courses not related to their core programs, therefore including them will predictably produce negative sentiments.

The messages contained phrases of local dialects, characters such as @, # and short forms of English words. However, most of the English terms that users replaced were stop words such as "de" for "the", "4" for "for", "2" for "to" and so on. A few messages had some important words required for the algorithms to produce good results replaced with local terms. For instance, "Ashock" was very frequent as a replacement to "I am shocked", "tnx" for "thanks", "siaa" for "fool", "ashawos" for "harlots", "mer3" or "tym" for "time", just to mention a few. This obviously is a challenge to any automated text mining algorithm. Therefore, these words were replaced with their equivalent standard English spellings.

Some of the messages contained jokes that did not convey the sender's true sentiments. Consider the following sentence which on first thought will seem to carry negative sentiment, but in actual fact, the sender was in good mood.

"I am very disappointed in this group and let me just warn you people about this, i do give you all the respect you deserve but there are certain things i won't tolerate from you and i want to serve this as a warning to you before it gets out of hand. How can you go around telling everybody that i said "Tomorrow" is PALM SUNDAY so i should learn. Haha. why are you scared...go on make others panic as i did to you now. Let me be the first to wish you HAPPY Palm sunday in advance."

On the other hand, it was noticed that students were frank in sharing their emotions through several of the messages such as:

"At long last, ma temperature is back to happiness",
"<<name>> dey gv pressure too much",
"<<name>> make school no de be me koraaa";

meaning the lecturer <<name>> was giving them too much pressure and <<name>> not making school an interesting place, apparently due to the perceived pressure. Cleaning and manual pre-processing of the data was done before applying the classification algorithms.

To avoid making faulty assumptions and conclusions as a result of the informal nature of messages in the corpus, two colleagues were tasked to perform inductive content analysis necessary to make the data ready for the application of classification algorithms. Their job was to clean the data and label a portion of the data for training. Since the aim was to make the sentiments author-centric, the researchers were requested to try as much as possible to impose sentiments by putting themselves in the position of the message sender and not just impose their own sentiments based on their emotions during labelling. In addition, author's subjective comments on others should be marked as subjective, since such comments may express approval or disapproval towards the source of the message or the message itself. Subjective messages about objects, people, issues, and places were also to be used. Each individual researcher gave labels to the same set of messages and handed over to the next researcher for evaluation. They finally came together to discuss areas where the labels they each gave to a given sentence was different. Messages that could not be classified as positive or negative were discarded since the classification is binary. Messages that had no bearing on education were eliminated. Most of them were religious. The training set was thoroughly reviewed resulting in the labelling of messages as positive or negative. In all, 69,373 instances were labelled, meaning that 563 messages were discarded based on the criteria described above. 70% of the labelled instances were used as training set and 30% as test set. Even though the annotators labeled the test set, the labels were removed before applying the classifier models. Most of the messages contained between 10 – 30 sentences. Those with more than 50 sentences were usually religious messages which were discarded.

2.1 Inter-rater Agreement

Since the task at hand is a single-label classification, Cohen Kappa (Cohen, 1960) was used to calculate the overall agreement between the labeling of the two researchers. In other words, the classification was such that no single data point could fall into more than one class since all classes must be mutually exclusive. The Kappa was used to measure the difference between the observed agreement and the expected agreement in the labeling. With a scale of -1 to 1, negative values indicate agreements less than chance, 0 being exactly what would be expected by chance and 1; almost perfect agreement. From the Concordance matrix shown in Table 2 (Sokolova & Bobicev, 2011), the kappa statistic was calculated for the researchers. The overall percent agreement was 0.94 with the Kappa being 0.85 between the annotators. A high Kappa does not mean the raters were correct about the classes they assigned but just an indication that there was a strong agreement between the labels they assigned.

$$kappa = \frac{\frac{a+d}{N} - \frac{f_1 g_1 + f_2 g_2}{N^2}}{1 - \frac{f_1 g_1 + f_2 g_2}{N^2}} \quad (1)$$

2.2 Feature Selection

Feature selection for classification involves the selection of attributes that are necessary for the classification task. For text classification, each unique word corresponds to a feature. The frequency of occurrence of this feature is its value. To limit the size of feature vectors, only unique words occurring more than ten times in the training set were considered as features. The data was preprocessed to improve the quality of selected features. First of all, characters such as @, #, *, -, $, + and hyperlinks which were not directly contributing to the meaning of messages were removed. In order to extract features from raw chats, tokenization was carried out to split the texts at non-letters so that anytime a non-letter was encountered the previous characters would denote a new token, therefore splitting the text into words. Since tokenization on non-letters alone may not produce the best outcome for sentiment polarity detection (i.e. some words on their own are meaningless), n-grams (n=2) were generated by removing tokens smaller than two and larger than fifty. A minimum token length of two was maintained to enable the algorithms to capture every smaller possible token that can be good determiners of emotions.

For emphasis, students relied on the use of repeated characters in words such as "*haaaaapppyyy*", "*haapyy*", "*yeeaaaah*", etc. in their messages. For such repeated characters, only two of the repeated letters in the word were kept. This was to enable coverage of all the various forms of such words in the corpus, reduce the dimensionality of the feature vectors and make the feature space less skewed. Stemming was carried out using the Porter stemmer (Porter, 1980) to reduce different forms of the same token to the same length. This enabled all the different forms of the word to be accounted for by the classification algorithms. For instance, the words "loved", "loving", "lovely", and "love" were reduced to the form "love". Common English words such as "a", "the", "it", "as", etc. usually referred to as stop words were also filtered out of the corpus. The case of all the tokens were transformed to lower case to allow for uniformity. To create the word vector, use was made of the term frequency–inverse document frequency (TF-IDF) (Kotu & Deshpande, 2015) method. It is a statistical weighing factor which determines how important a word is to a document. How frequent a term occurs in a document defines its term frequency, with the inverse document frequency ensuring that the effect of more common words in a document is controlled. If we let n_k be the number of times a keyword k appears in a document, and n be the number of terms in the document, then TF is given by

$$TF = \frac{n_k}{n} \tag{2}$$

Letting N = the number of documents under consideration, and N_k = the number of documents in which the term k appears, IDF can be obtained by:

$$IDF = \log_2\left(\frac{N}{N_k}\right) \tag{3}$$

Finally, adding the term frequency, we obtain:

$$TF - IDF = \frac{n_k}{n} * \log_2\left(\frac{N}{N_k}\right) \tag{4}$$

2.3 Classification of Student's Messages

Three individual classifiers were trained and tested after which they were used together in an ensemble classifier based on majority voting (Rokach, 2010). The individual classifiers were Naïve Bayes, Support Vector Machines and Decision Trees. 70% of the data set was used to train the algorithms (using cross-validation). 30% was then reserved to test the models. To obtain accurate classification results, it was important to train the classifiers with balanced set. Naïve Bayes is simple and accurate but can be shown to produce inaccurate results for polarity detection if the training set does not contain a balanced number of positive and negative data (Wan & Gao, 2015).
1) *Naïve Bayes classification*: A Naïve Bayes (NB) classifier was trained using carefully labelled data from the dataset. With its strong (naïve) independence assumption and simplicity, it is known to perform very well on text classification task (Mccallum & Nigam, 1998), (Alsubaey, Asadi, & Makatsoris, 2015). This probabilistic classifier assumes word independence and does not allow the conditional probability of the presence/absence of a word to affect the conditional probability of the presence/absence of other words in the document. It has the advantage of being able to use a small amount of training data to estimate the means and variances of the variables used for the classification task (Crc & Hofmann, 2014). The concept of NB classification is based on allowing every single feature to contribute in the determination of a class (label) for each input text (Steven, Ewan, & Loper, 2009). The frequency of each label in the training set is used to calculate its prior probability. To obtain the likelihood estimate for the label, its prior probability is aggregated with those generated from all the features after which the label with the highest likelihood estimate is assigned to the input. Assuming l is a label and f features,

then the conditional probability that a given input will be assigned to the label *l* given that its features are known is given by:

$$P(l \mid f) = \frac{P(l \cap f)}{P(f)} \tag{5}$$

Since *P(f)* will be the same for every label, we can apply total probability rules to the denominator of equation (5). The resulting expression then becomes:

$$P(l \mid f) = \frac{P(l) * P(f \mid l)}{\sum_{l \in ls} P(l) * P(f \mid l)} \tag{6}$$

where *ls* is the set of labels used for the classification task.

To avoid the high influence of zero probabilities, the Laplace correction was applied. This algorithm assumes a large training set such that adding one to each count is assumed to be negligible in terms of the estimated probabilities which at the same time prevents zero probability values. To estimate how the NB algorithm will perform on the training set, Cross-Validation (X-Validation) was applied. X-Validation partitions the training set into k subsets of equal sizes out of which one subset is kept for testing. In other words, the inputs to the testing sub-process, plus the remaining k−1 subsets are used as training datasets. X-Validation is repeated k times such that, each of the k subsets is used exactly once as the test data. The k results from the k iterations are then averaged to produce a single estimation. X-Validation predicts the fit of a classifier to a hypothetical dataset in the absence of separate test set to help avoid over fitting; a scenario where a learning algorithm is perfectly optimized and fits its training data very well but fails to perform well on some independent training set. In this case, the value of k was set to 10 for the NB classifier. Stratified Sampling technique was used to ensure that the class distribution in the k subsets was the same as in the entire training set. The NB classifier obtained 0.71 accuracy, 0.70 precision, 0.85 recall and an F-score of 0.77.

2) *Support Vector Machines (SVM) Classification*: SVMs take a different approach to classifying texts as opposed to NB classifiers. Their ability to learn does not depend on the dimensionality of the feature space but based on the principle of searching for a separating hyper-plane between different classes; in this case and the case of binary classification the classes are two - positive and negative. The points closest to the separating hyper-plane are called Support Vectors. Their placement is such that they must be as far as possible from the separating hyper-plane. In other words, SVM is a solution to an optimization problem that will choose a hyper-plane with the maximum margin among many possible boundaries separating the classes. This is important because the larger the margin, the lower the generalization error of the classifier. Trained with labeled examples, SVM is able to assign new examples into one class or the other. They have the ability to support high dimensional input spaces, use over-fitting protection, and treat text classification as linearly separable classification problems (Joachims, 1998). In addition, they perform very well on sparse document vectors (vectors with few entries with the rest of the entries being zero) such as those generated for purposes of text classification. SVMs are excellent learners even in their simple form as they learn with linear threshold function. Theoretically, SVM's training set can be represented as (Sassano, 2003),

$$(x_i y_i),\ldots,(x_m y_m) x \in \Re^n, y_i \in \{positive, negative\}$$

Letting K represent the kernel function, $b \in \mathbb{R}$ the threshold, and $\propto_i$ the weights, we can define the decision function g as

$$g(x) = \operatorname{sgn}(f(x)) \tag{7}$$

with

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x) + b \tag{8}$$

The weights $\propto_i$ must satisfy the following constraints; given that C is the cost of misclassification.

$$\forall_i : 0 \le \alpha_i \le C, and \sum_{i=1}^{m} y_i \alpha_i = 0$$

The training vectors (x_i) with non-zero weights ($\propto_i$) are called the support vectors. For linear SVMs (as in this case), the kernel function K is defined as the dot product

$$K(x_i, x) = (x_i, x)$$

(9)

with the decision function becoming

$$f(x) = w.x + b \tag{10}$$

where

$$w = \sum_{i=1}^{m} y_i \alpha_i x_i \tag{11}$$

Training an SVM involves finding solution to $\propto_i$ and b by optimizing the problem below:

$$\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{12}$$

subject to the same constraints on the weights $\propto_i$.

We leveraged on the effectiveness of SVM and the Linear separable property of texts to classify each input x_i from the test set into one of two classes; positive or negative. The Support Vector Machine (Linear) in RapidMiner was trained with labeled data. This version of SVM is a non-probabilistic classifier that is implemented in java as *mySVM*. It was chosen because it is a fast algorithm and has the ability to provide efficient results for many learning tasks. To avoid over-fitting and over-generalization, a complexity constant of 0.001(tolerance for misclassification) was used which is neither too large to cause over-fitting nor too low to over-generalize. Ten-fold Cross-Validation was used to determine how well SVM will perform on the training data and we achieved *0.75* accuracy, *0.76* precision, *0.82* recall with an F-score of *0.79*.

3) Decision Trees Classification: Decision trees are rule-based classifiers that do not use distance measures for classification. The nodes of the tree serve as decision points. The value of the feature of a node is used as criteria to select the next node. The leaves represent the class in which an instance is categorized. When applying a decision tree for binary classification, each feature in the training set is considered separately. Based on the value of the current feature, the training set is divided into instances of two subsets and the feature with the purest subset chosen. The value of this attribute is then set as the decision condition by the current node. Since the aim was to measure the gain in purity from parent to children down the hierarchy of the tree, gain ratio was selected as the criterion on which attributes will be selected for splitting. The gain ratio (Δ) is given by (Janert, 2011)

$$\Delta = I(parent) - \sum_{children,j} \frac{N_j}{N} I(child, j) \tag{13}$$

where I is the purity (or impurity) of a node, N_j is the number of elements assigned to child node j, and N is the total number of elements at the parent node. The objective is to find a splitting that will maximize the gain ratio. Decision trees are fast to build, performs well in the presence of noise and large data. Percentual pruning was used to avoid leaf nodes having few elements in an attempt to prevent over-fitting. Percentual pruning is a method which ignores words according to their percentage of appearance in all documents. This is achieved by preventing the algorithm from continually splitting the training set until a point where all leaf nodes are completely pure. Ten-fold Cross-Validation was used to determine how well the Decision Tree will perform on the training data as described in earlier sections, and it achieved 0.70 accuracy, 0.74 precision, 0.75 recall with an F-score of 0.75.

4) Ensemble Classification: Experimenting with the above three classifiers showed that some of the messages were classified differently depending on which classifier was under consideration. In an attempt to combine the strengths of these classifiers, a decision was taken to combine them in an ensemble which will make use of simple majority voting. The ensemble classifier was trained with the same training set as the individual classifiers, and applied voting to assign the majority votes of all predicted values to an unknown example. For classification tasks, all the individual classifiers making up the ensemble receives the training set to generate their individual classification models. The voting process takes place by considering the votes of all the individual classification models and assigning the predicted class with maximum votes to the unlabelled instance. It uses the predictions of the base learners to make a combined prediction by simple majority voting; that is the one most often predicted by the different classifiers. If the k^{th} classifier produces $y_k(x)$ as its classification and g(y.c) is an indicator function, then:

$$class(x) = \underset{c_i i \in dom(y)}{\arg\max} \left(\sum_k g(y_k(x), c_i) \right) \tag{14}$$

where

$$g(x,c)=\begin{cases}1, y=c\\0, y\neq c\end{cases} \tag{15}$$

Another version of this classifier was also implemented by relying on the confidence or probability scores (Asker & Maclin, 1997), (Fung, Yu, Wang , Cheung, & Liu, 2006)to build the ensemble. Comparing the performances of the two ensemble classifiers, the later underperformed so a decision was taken to use the former for further analysis in this paper. Split Validation was used. This was to enable the model randomly split up the data into training and test sets and then do evaluation to estimate how the performance of the learner will be in practice. Significantly, the ensemble classifier outperformed the three individual classifiers. An accuracy of 0.76, precision of 0.72, recall of 0.92 and F-score of 0.80 were obtained. This implementation was therefore used for further analysis.

5) Evaluation Measures for Text Classification: Measures such as Precision, recall, accuracy and F-score are usually used to show the performance of classification algorithms. These measures can be evaluated from the confusion matrix shown in Table 2.

Precision is a measure of the fraction of the results from the dataset that are classified correctly.

$$precision(P)=\frac{t_p}{t_p+f_p} \tag{16}$$

Recall measures the fraction of correct items in a class that the algorithm actually classifies as belonging to that class.

$$\mathrm{Re}call(R)=\frac{t_p}{t_p+f_n} \tag{17}$$

Accuracy is a measure that finds the ratio of the "*true*" values and the total instances that occur in the dataset.

$$Accuracy(A)=\frac{t_p+t_n}{t_p+t_n+f_p+f_n} \tag{18}$$

F-score is the harmonic mean of precision and recall. It is the weighted average of precision and recall. The best and highest value for this measure is 1 with 0 being the least and worst value.

$$F-score=2*\frac{P*R}{P+R} \tag{19}$$

3. Results and Discussion

Table 3 shows the comparison of the performance parameters of the classification algorithms. On average, the ensemble classifier performed better in predicting the polarity of sentiments expressed in the messages than the three individual algorithms.

Using the supervised ensemble classifier, unseen data was classified for each option in Table 1. Fig. 3 shows the percentage of messages classified as having either negative or positive polarity per option. The overall sentiment polarities for *BSc. Computing with Accounting level 400*, BSc. Computer Science and Diploma in Computer Science were negative as opposed to *BSc. Computing with Accounting level 300* and BSc. Information Technology which were positive. Significant is the fact that, the polarity of sentiments for all final year classes except IT were negative, whiles lower level class (*BSc. Computing with Accounting level 300*) were positive. To understand this scenario, the distribution of words, phrases and/or activities per option that could serve as indicators to students' emotions in relation to their studies were considered. For example, the phrase *examinations period ("exam_period")* for BSc. Computing with Accounting Level 400 was normally distributed and negative (Fig. 4) with high density as compared to its use as positive by the same class. There could be several reasons for the high-density negativity associated with this phrase among which includes the anxiety and pressure that accompanies examinations. Other activities such as *student registration ("regist_student")* and *learning ("learn")* were associated with negative sentiments for the same class. However, the *end of a lecture ("lectur_end")* as shown in Fig. 5, getting a *past question ("pasco")*, and the anticipation of a possible *leakage* of a question paper *("leak")*, were found to be associated with positive sentiments for CSC, CwA400 and Diploma. For negative sentiments, some of the top indicative words for the BSc Computer Science class were *"sleep"* and *"assignment"*. In other words, it is possible that students were loaded with several assignments causing sleep problems hence leading to the outpouring of negative sentiments. In addition, weekends ("*weekend*") were no longer happy

moments due to the numerous assignments.

For BSc Information Technology, the availability of *"lecture handout"* had positive sentiments as opposed to the taking of *"lecture notes"* in class. There was also a favorable attitude towards ongoing courses such as Java for CSC, Visual Basic and Computer Hardware Architecture for CwA300.

Table 5 shows the list of words, names and phrases serving as direct basis for students' emotions and sentiments towards the learning process.

Aside the determination of whether a message has positive or negative polarity, the "not so simple" form of sentiment analysis which may involve the detection of the six Ekman emotions (Ekman, 2004) such as joy, anger, sadness, fear, surprise and disgust was also carried out. This is achievable through the use of sentiment analysis together with effective visualization [42], (Rashid, Aitken, & Fels, 2006), (Subasic & Huettner, 2001). Figs. 6(a)-(e) shows how the use of the six emotional words varied among the five options over the period.

A system was developed to identify and calculate the percentages of each of the six emotional words by taking the ratios of the number of occurrence of a given emotional word to the total number of emotional words in the messages. A given word was identified as an emotional word based on the approach used by (Mohammad, 2011).

The emotional word "*joy*" was used much more on average by BSc Computing with Accounting 300 and BSc Information Technology than any other of the six emotional words. For BSc Information Technology, this could partly be attributed to the size of the class; perhaps lecturers had time for individual students since they were only 18 in class as compared to 200 students in BSc Computer Science. Therefore, they may have little difficulty in organising themselves to understand what a lecturer is teaching. As stated earlier, BSc Computing with Accounting 300 had a favourable attitude towards two of their core courses, namely Visual Basic and Computer Hardware Architecture, this could be one of the reasons why the class seemed to be in a "*joyous mood"* for most of the period under consideration. The trend of the emotional words in Fig. 6 could be as a result of several reasons. As to the specific reasons, we can only speculate. However, it is important to notice that Fig. 3 and Fig. 6 produced similar outcomes.

3.1 Conclusion

In this paper, supervised ensemble classification was applied on WhatsApp group messages to predict the mood of students towards learning in an academic setting. An F-score of 0.80 was obtained for the classification, the outcome was then used to identify some possible reasons for the sentiments students expressed. Problems discovered included lack of sleep due to heavy load of assignments, lack of accommodation or inability to afford one, pressure of examination period, frustration during course registration, and difficulty in grasping the concepts in some of the courses.

This approach can serve as a guiding tool for educational policy makers to make informed decisions in order to incorporate students' concerns based on the problems and emotions so discovered. The advantage is that, both students and lecturers will benefit since lecturers will identify the "academic problems" of students in their class in a timely manner to enable them refine their style of teaching or redesign their teaching methods.

However, it is not every student who has a WhatsApp phone and the challenge as to whether students are truthful about their situation at the time of posting messages or are relying on parables or figure of speech is still paramount. Privacy issues are acknowledged and addressed such that the identities of students were to be kept anonymous. In addition, the scope of work in this paper does not cover analysis of multimedia content of the messages.

As future work, this analysis will be carried to the level of individual students. This will enable us to relate student's sentiments to their academic performance retrievable from the University's student management information system (MIS) shown in Fig. 1.

References

[1] Alsubaey, M., Asadi, A., & Makatsoris, H. (2015). A Naïve Bayes Approach for EWS Detection by Text Mining of Unstructured Data: A Construction Project Case. SAI Intelligent Systems Conference, 164-168.

[2] Arroyo, I., Woolf, B., Burelson, W., Muldner, K., Rai, D., & Minghui, T. (2014). A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. Int J Artif Intell Educ, 24, 387-426.

[3] Asker, L., & Maclin, R. (1997). Ensembles as a Sequence of Classifiers. Proc. of IJCAI, 860-865.

[4] Carchiolo, V., A. L. B, & Malgeri, M. (2015). Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics. ITBAM, LNCS 9267, 16-24.

[5] Chen, X., Member, S., Vorvoreanu, M., & Madhavan, K. (2014). Mining Social Media Data for Understanding Students' Learning Experiences. IEEE Trans. Learn. Technol, vol. 7(no. 3), 246-259.

[6] Chikersal, P., Poria, S., Cambria, E., Gelbukh, A., & Siong, C. (2015). Modelling Public Sentiment in Twitter: Using Linguistic Patterns to Enhance Supervised Learning. Part II, LNCS 9042, 49–65.

[7] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas., 37–46.
[8] Crc, H., & Hofmann, M. (2014). RAPID MINER Data Mining Use Cases and Business Analytics Applications. Taylor & Francis Group, CRC Press.
[9] Dorwal, P., Sachdev, R., Gautam, D., & Jain, D. (2016). Role of WhatsApp Messenger in the Laboratory Management System: A Boon to Communication. J Med Syst, 3-7.
[10] Ekman, P. (2004). Emotions revealed (Vol. vol. 12).
[11] Fung, G., Yu, J., Wang, H., Cheung, D., & Liu, H. (2006). A balanced ensemble approach to weighting classifiers for text classification. IEEE International Conference on Data Mining. no. 4, pp. 869–873. ICDM.
[12] Gowda, S., Baker, R., Corbett, A., & Rossi, L. (2013). Towards Automatically Detecting Whether Student Learning is Shallow. Int J Artif Intell Educ, 50-70.
[13] Guler, C. (2016). Use of WhatsApp in Higher Education: What' s Up with Assessing Peers Anonymously? J. Educ. Comput., 1-18.
[14] Janert, K. (2011). Data Analysis with Open Source Tools.
[15] Janssen, D., Tummel, C., Jeschke, S., & Richert, A. (2015). Sentiment Analysis of Social Media for Evaluating Universities. Proceedings of Second International Conference on Digital Information Processing, Data Mining, and Wireless Communications, 49-62.
[16] Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of ECML-98, 10th European Conference on Machine Learning, (pp. 137-142).
[17] Kalra, A., & Karahalios, K. (n.d.). TextTone: Expressing emotion through text.
[18] Khanna, V., & Sambandam, S. (2015). "WhatsApp" ening in orthopedic care: a concise report from a 300-bedded tertiary care teaching center. Eur J Orthop Surg Traumatol.
[19] Korenek, P., & Marián, Š. (2013). Sentiment analysis on microblog utilizing appraisal theory. Springer Sci. Media New York.
[20] Kotu, V., & Deshpande, B. (2015). Predictive Analytics and Data Mining Concepts and Practice with RapidMiner. Elsevier Inc.
[21] Mccallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 workshop on learning for text categorization.
[22] Mensa-Bonsu, M. (2016, March). The old order changeth and giveth way to the new: WhatsApp debuts as a means of substituted service in Ghana. Oxford Univ. Commonw. Law J. ISSN, vol. 9342.
[23] Mensah, P. (2015). Textual prediction of attitudes towards mental health. Int. J. Knowl. Eng. Data Min., 3(3/4), 274-285.
[24] Merceron, A., & Yacef, K. (2003). A Web-Based Tutoring Tool with Mining Facilities to Improve Learning and Teaching. Artif. Intell. Educ, 201-208.
[25] Mohammad, S. (2011). Sentiment Analysis of Mail and Books.
[26] Ortigosa, A., Martín, J. M., & Carro, R. M. (2013). Sentiment analysis in Facebook and its application to e-learning. Comput. Human Behav.
[27] Porter, M. (1980). An algorithm for suffix stripping. Program, 130-137.
[28] Rashid, R., Aitken, J., & Fels, D. (2006). Expressing Emotions Using Animated Text Captions. Lncs, 4061, 24-31.
[29] Rokach, L. (2010). Ensemble-based classifiers. Artif Intell Rev, 1-39.
[30] Sabourin, J., Shores, R., Mott, B., & Lester, J. (2013). Understanding and Predicting Student Self-Regulated Learning Strategies in Game-Based Learning Environments. Int J Artif Intell Educ, 23, 94-114.
[31] San Pedro, O., d Baker, R., & Rodrigo, M. (2014). Carelessness and Affect in an Intelligent Tutoring System for Mathematics. Int J Artif Intell Educ, 24, 189-210.
[32] Sassano, M. (2003). Virtual Examples for Text Classification with Support Vector Machines. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, (pp. 208-215).
[33] Seufert, M., Hoßfeld, T., Schwin, A., Burger, V., & Tran-gia, P. (2016). Group-based Communication in WhatsApp. IoP Workshop co-located with Networking 2016, 536–541.
[34] Sokolova, M., & Bobicev, V. (2011, September). Sentiments and Opinions in Health-related Web Messages. Proceedings of Recent Advances in Natural Language Processing, 132-139.
[35] Steven, B., Ewan, K., & Loper, E. (2009). Natural Language Processing with Python.
[36] Subasic, P., & Huettner, A. (2001). Fuzzy Semantic Typing. IEEE Trans. FUZZY Syst, vol 9(no. 4), 483-496.
Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics. (2015). ITBAM, LNCS 9267, 16–24.
[37] Wai, D., Chu, K., Wai, P., & Lam, M. (2015). Analysis of Student Behaviors in Using WeChat /WhatsApp for Language Learning at Diploma Level in Hong Kong: A Pilot Test. International Symposium on Educational Technology Analysis, 106-110.
[38] Walker, E., Rummel, N., & Koedinger, K. (2014). Adaptive Intelligent Support to Improve Peer Tutoring in Algebra. Int J Artif Intell Educ, 24, 33-61.

[39] Wan, Y., & Gao, Q. (2015). An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. IEEE 15th International Conference on Data Mining Workshops, 1318–1325.
[40] Westerfield, G., Mitrovic, A., & Billinghurst, M. (2015). Intelligent Augmented Reality Training for Motherboard Assembly. Int J Artif Intell Educ, vol. 25, 157-172.
[41] Yeboah, J., & Ewur, D. (2014). The Impact of Whatsapp Messenger Usage on Students Performance in Tertiary Institutions in Ghana. J. Educ. Pract, vol. 5(no. 6), 157-164.
[42] Younis, E. (2015). Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools : An Empirical Study. Int. J. Comput. Appl., 112(no.5), 44-48.
[43] Zorrilla, M., Álvarez, E., & García-Saiz, D. (2015). A parametrisable method for measuring online attendance in e-learning tools. Int. J. Technol. Enhanc. Learn., vol 7(no. 4), 289-308.

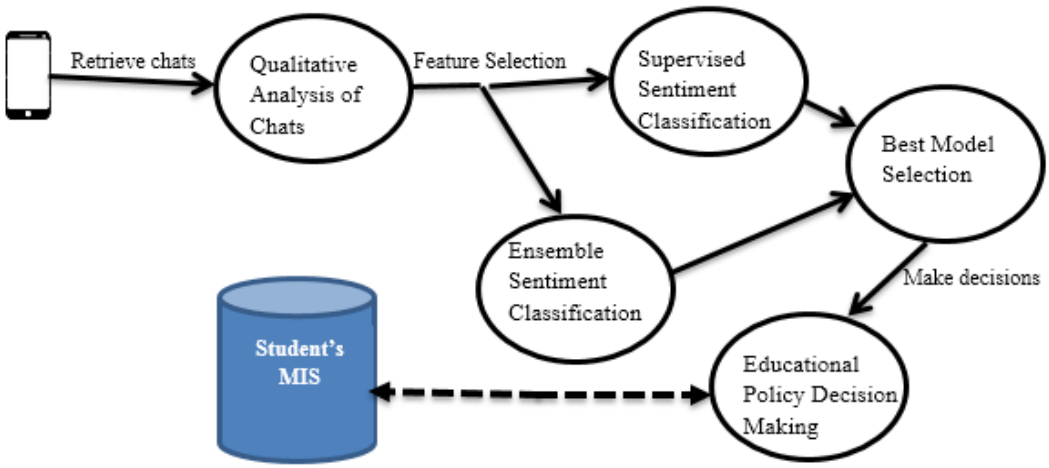

Fig. 1 The workflow adopted in this paper.

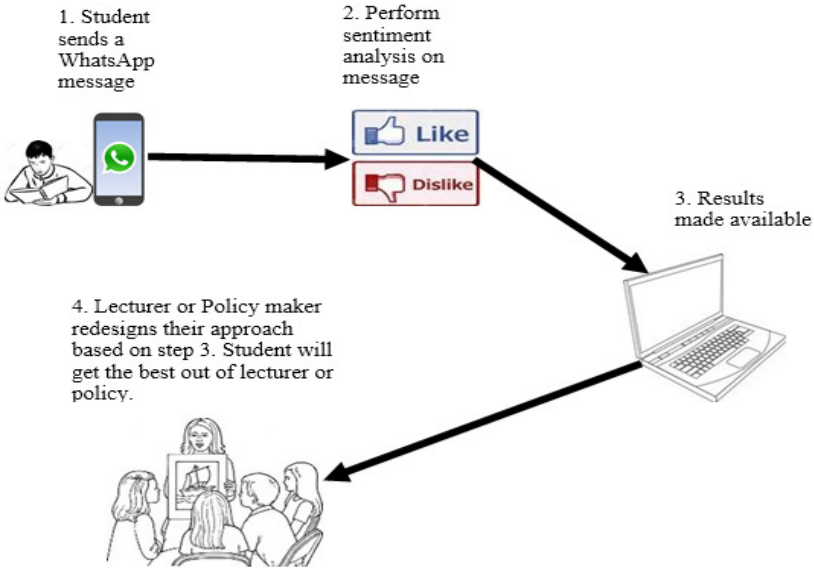

Fig. 2 One important practical use of sentiment analysis in education.

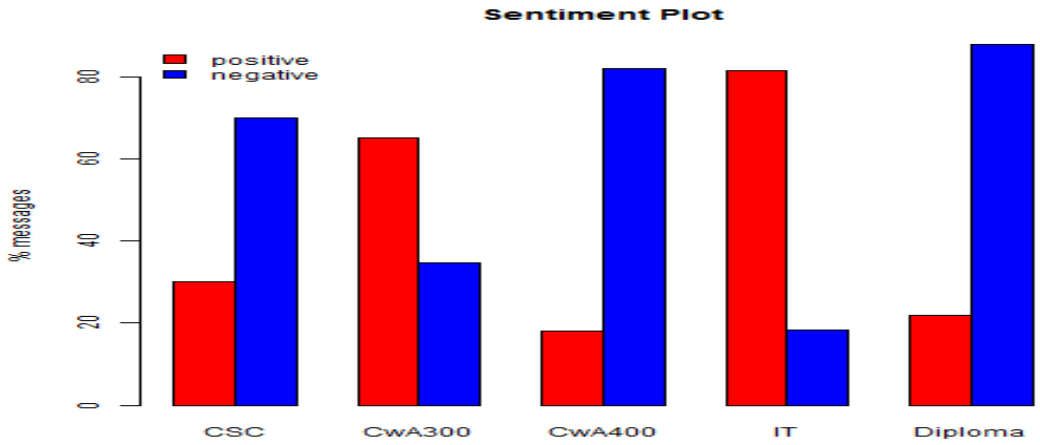

Fig. 3 A bar graph of the classified messages showing the percentage of messages falling in each category for each program option; that is CSC- BSc. Computer Science, CwA300- BSc. Computing with Accounting Level 300, CwA400- BSc. Computing with Accounting Level 400, IT-BSc. Information Technology, and Year Two Diploma in Computer Science.

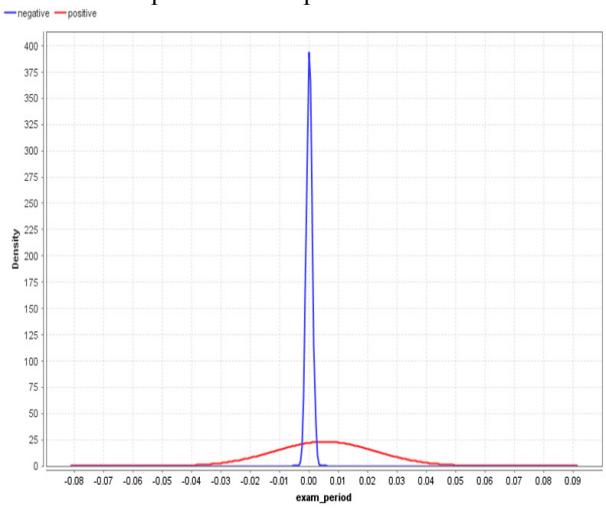

Fig. 4 Probability density functions (pdf) for the phrase "*exam_period"* for BSc. Computing with Accounting Level 400. The red line shows the distribution of the phrase when its use is associated with positive sentiments, blue shows the distribution when its use is associated with negative sentiments.

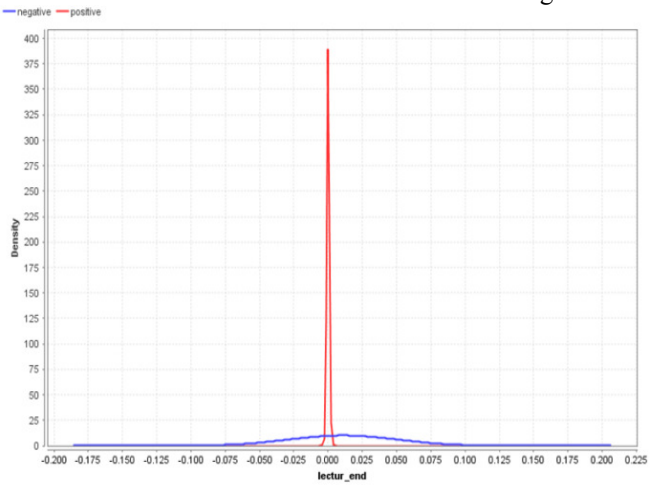

Fig. 5 Probability density functions (pdf) for the phrase "*lecture_end"* for BSc. Computer Science. The red line shows the distribution of the phrase when its use is associated with positive sentiments, blue shows the distribution when its use is associated with negative sentiments.

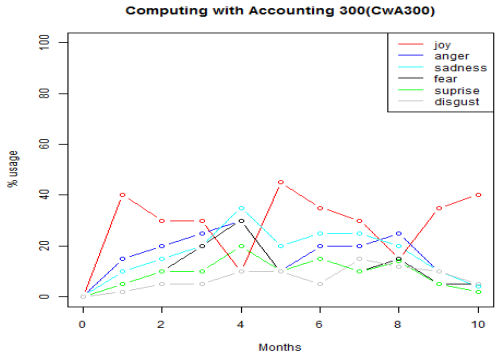

Fig. 6 (a)

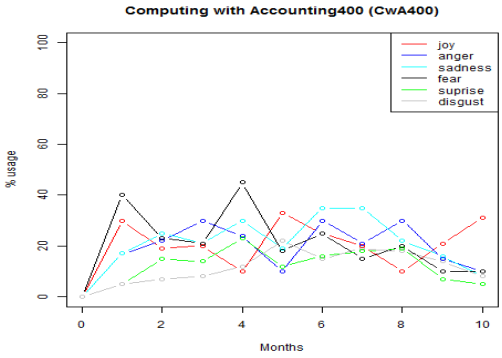

Fig. 6 (b)

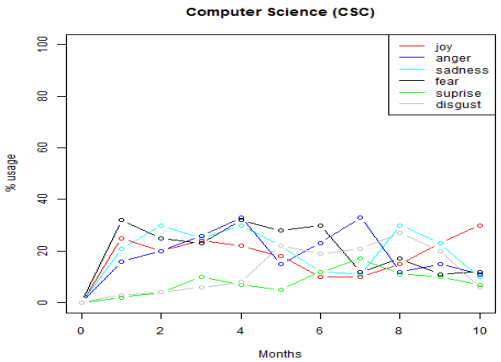

Fig. 6 (c)

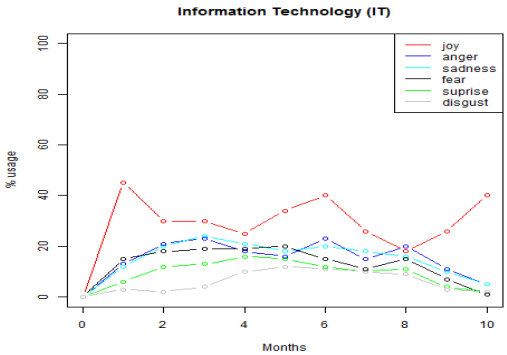

Fig. 6 (d)

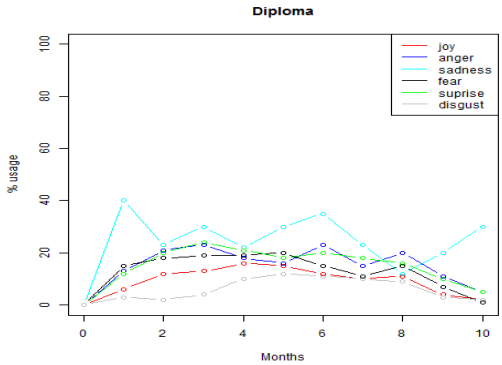

Fig. 6 (e)

Fig. 6 The use of emotional words in the messages for all the options (a) – (e). On the x-axis, numeric codes are used to represent the months during which the messages were sent. That is 1=September, 2=October, 3=November, 4=December, 5=January, 6=February, 7=March, 8=April, 9=May and 10=June.

TABLE I
NUMBER (#) OF MESSAGES PER OPTION. A SINGLE MESSAGE MAY CONTAIN MORE THAN 10 SENTENCES.

Program Option	**#Students**	**#Raw Messages**	**#Cleaned Messages**
Computer Science	76	12,332	12,125
Computing with Accounting 300	112	34,834	34,733
Computing with Accounting 400	72	20,371	20,150
Information Technology	18	1,839	1,829
Diploma in Computer Science	25	560	536
Total	**303**	**69,936**	**69,373**

TABLE 2
CONCORDANCE MATRIX.

2nd observer	***1st observer***		
	YES	***NO***	***Total***
YES	*a*	*b*	g_1
NO	*c*	*d*	g_2
Total	f_1	f_2	*N*

TABLE 3
CONFUSION MATRIX

	Predicted	
True	***positive***	***negative***
positive	*true positive* (t_p)	*false negative* (f_n)
negative	*false positive* (f_p)	*true negative* (t_n)

TABLE 4
PERFORMANCE COMPARISON OF THE CLASSIFIERS

Algorithm	**Precision**	**Recall**	**F-score**	**Accuracy**
SVM	0.76	0.82	0.79	0.75
NB	0.70	0.85	0.77	0.71
D. TREE	0.74	0.75	0.75	0.70
ENSEMBLE	0.72	0.92	0.80	0.76

TABLE 5
TOP WORDS, PROBLEMS, ACTIVITIES, NAMES AND PHRASES ASSOCIATED WITH STUDENTS' SENTIMENTS

Sentiment Polarity	***Top Indicative Words or Phrases***
Negative	*accommodation, book, campus hostel, hot temperature, student affairs, student finances, student loan, class, examinations office, exams period, learn, lecture note, lecture time, results, wifi, register students, sleep, assignment, work, library, weekend,*
Positive	*Android, affairs of src, student forum, student leader, course, java, web design, visual basic, hardware architecture, lecture handout, leakage, uew, ucc, ucc src, legon, pasco, teach, lecture end*