



Ensemble Classification of Sentiments for Monitoring Students' Academic Progress

Mensah Kwabena Patrick¹, Stephen Akobre²

¹Department of Computer Science, P.O.Box 24, UDS, Navrongo, Ghana.

²Department of Computer Science, P.O. Box 24, UDS, 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 emotions to their peers without fear. To obtain immediate feedback on problems hindering academic progress, we applied supervised learning algorithms to classify sentiments in WhatsApp group chats of University students. An ensemble classifier made up of Naïve Bayes, Support Vector Machines, and Decision Trees outperformed the individual classifiers and produced an accuracy of 0.76, 0.92 recall, a precision of 0.72 and 0.80 F-score, an indication that we can accurately approximate students' emotions from their private chats. This paper then proposes a method to determine the relationship between students' sentiments and their academic performance with the objective of creating an avenue for educational authorities to cost effectively monitor issues hindering students' academic progress.

Keywords: WhatsApp, Ensemble Classification, Sentiments, Education, Academic Progress, Naïve Bayes.

I. Introduction

Many of the available research on sentiment analysis (Chen et al. 2014); (Chikersal et al. 2015); (Carchiolo et al. 2015) uses data from the major social media sites such as Twitter, Facebook and web-content such as blogs. Researchers believe that users have the freedom to freely express their concerns and opinions on social media without fear of victimization. However, in practice user's conversation on such sites is still accessible to a wide range of audience defeating the privacy issue. On the other hand, messaging platforms such as WhatsApp allows users to form groups. Messages to these groups can be sent and read only by members of the group, making it the appropriate medium for users to express their true emotions. As a result, University students in our part of the world 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. These messages are an enormous source of data from which we can gain insight into the learning process. On the other hand, most educational data mining research focuses on either the use of online learning data, course management systems data or classroom technology data. A few concentrates on the use of Social Media Mining (Abdelrazeq et al. 2015) to understand student's sentiments. Very few make the attempt to associate these sentiments to individual student's academic progress. To the best of our knowledge, there is no research that uses unstructured mobile data (messages) from student WhatsApp groups to mine the sentiments and opinions of students leading to the discovery of the relationship between students' sentiments and their academic progress.

The problems this paper attempts to solve are in twofold; First, students encounter different type of problems which they find difficult to discuss with adults but less difficult to discuss informally with their colleagues. How can we get an idea about the core problems facing them? A solution to this problem will be beneficial to educational authorities who wants to formulate strategies for timely intervention. Secondly, can we establish a relationship between sentiments of students derived from these problems and their academic performance? In so doing, we will be providing concrete areas for educational authorities to act on.

In a broader sense, this paper seeks to achieve the following objectives:

1. Identify via supervised learning the polarity of sentiments derived from emotions and major problems hindering students' academic progress.
2. Use Association Rule Mining to determine the existence of a relationship between student's sentiments and their academic progress. In other words, can their participation in the group chats be a good predictor of their academic performance?

Our motivation strongly hinges on the fact that students are more comfortable discussing their emotions, problems, experiences and likes in an informal way with their friends on a private messaging platform like WhatsApp than on public social network. About 70% of students in our part of the world own a smart phone with WhatsApp

installed and in use. They spend countless hours on their phones sending and receiving WhatsApp messages, making it the most used messaging platform in many developing countries (Jisha & Jebakumar 2014). 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. With our aims and motivation clearly defined, we collected WhatsApp messages from four program options; BSc. Computer Science, BSc. Information Technology, BSc. Computing with Accounting and Diploma in Computer Science.

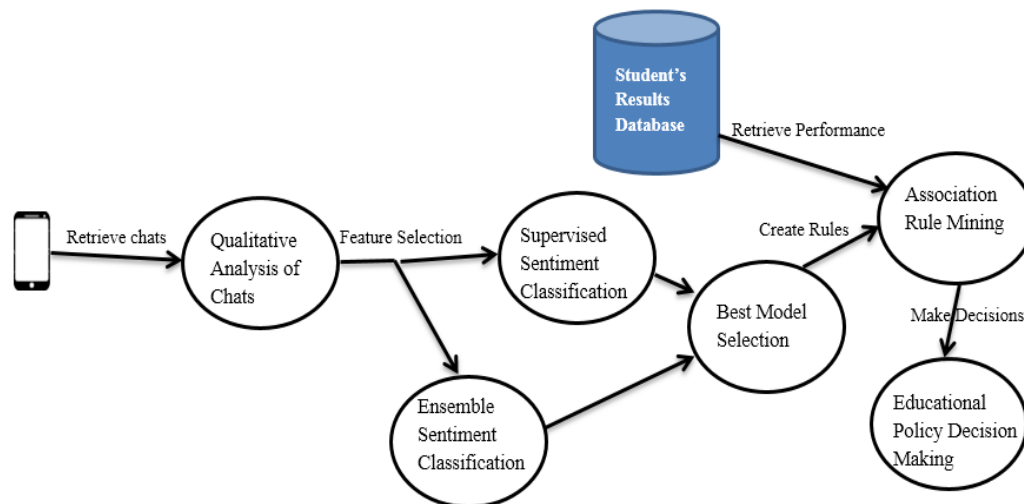


Fig. 1 The workflow adopted by this paper

We collected group messages for each of these options between September 2016 to June 2017, which is one complete academic year. Table 1 shows the total number of messages retrieved from each program option over the said period. Three binary classifiers and an Ensemble learning algorithm based on majority voting were trained and used for the classification of the polarity of group and individual messages. Comparatively, the Ensemble classifier outperformed the individual classifiers and was used for further analysis. The workflow adopted in this paper is depicted in Fig. 1.

We summarize the contributions of this paper as follows:

1. A demonstration that students' problems and sentiments can be understood by mining cheaply sourced private messages to obtain better results as opposed to mining social media or course management systems data.
2. It proposes a framework for determining the relationship between the polarity of student's sentiments obtained from their chats and their academic performance.
3. To serve as a tool for identifying troubled students at the early stage.

II. Related Work

2.1 WhatsApp as a means of Secure Private Communication

The success of Data mining algorithms largely depends on the availability of hidden patterns in large collections of data. An algorithm may perform well during training, but will fail to detect any meaningful pattern if the area of application is devoid of target patterns. To this end, we took into consideration the Goffman's theory of social performance (Goffman 1959), which demonstrates that the relaxing environment of backstage during face-to-face interactions results in unplanned actions (responses), before selecting WhatsApp as our source data. Although there are security concerns associated with the use of WhatsApp¹, research shows that students are becoming glued to it despite its numerous negative effects on academic performance (Yeboah & Ewur 2014). The challenge as to whether users are being frank about their situation at the time of posting messages on social media or they are relying on parables is still paramount. For instance, our analysis on Facebook comments of University students indicated the frequent use of the sentence "*happily sleeping*". On the other hand, the phrase "*no no how to teach, sleeping*" was used on WhatsApp instead of the former to represent the same subject. This turned out to be a jargon showing that a lecturer doesn't teach to their understanding, as a result of which most of them fall asleep in class. In other instances, the phrase "*father Christmas*" came up frequently in their Facebook comments. In

¹ Schrittwieser, S. et al., Guess Who's Texting You? Evaluating the Security of Smartphone Messaging Applications. Available at: https://www.internetsociety.org/sites/default/files/07_1.pdf

their WhatsApp group chats, they preferred to use the phrase “*he will surely give areas*”. In both cases, it is apparent that students were clearer and straight to the point in their WhatsApp messages as opposed to the Facebook comments. This could be attributed to the fear of victimization should they be more open on Facebook which is an open social media platform.

2.2 Mining the Web and Social Media Data

Social media data has been used by researchers to perform different data mining tasks. Application areas include mining to improve organizational image (Roosevelt 2012); (Boom 2015), forecasting of events such as disease or disaster outbreaks and their intensity² (Ashktorab et al. 2014). Another important application area includes mining both the structure and data³ of social media for conflict prevention and resolution (Mensah & Akobre 2015). It has been demonstrated that messages from Social Media sites such as Twitter can be harvested and mined to understand the experiences users of such sites go through (Chen et al. 2014). The attitudes of authors and other stakeholders have been predicted from publicly available web data such as news items and article publications (Mensah 2015). Research in this category usually uses a mix of the techniques in machine learning, data mining and natural language processing. Binary classification and sentiment analysis makes up a chunk of the work already done in this field with others using multi-class (Chen et al. 2014) classification. Whichever approach is deemed suitable depends on the task at hand. Algorithms such as Naïve Bayes, Decision trees, Support Vector Machines, Logistic Regression and Maximum Entropy are mostly applied after feature selection. Majority of the research uses data from Twitter to predict depression (Choudhury et al. 2013), predict Tie Strength (Gilbert & Karahalios 2009), categorize Movie Reviews (Harer & Kadam 2014), sentiment analysis (Kiritchenko et al. 2014); (Moreira et al. n.d.); (Deng et al. 2014); (Gamallo & Garcia 2014) in health related topics (Sokolova & Bobicev 2011); (Ali et al. 2013) and Natural Language processing tasks such as text mining (Yano et al. 2012).

2.2.1 Mining Data from Private Messages

Private messages and chats are likely to convey accurate information on the parties communicating. The potential for identity misrepresentation is far less risky than in mainstream social media. As a result, mining private messages has been used to successfully identify people who may cause harm to themselves (Matykiewicz et al. 2009); (Pestian et al. 2008). Application of Sentiment analysis on such messages can predict emotions which subsequently can be used to identify the gender of the sender, potentially leading to the development of tools and services that would monitor individual communications for evaluation and assessment (Mohammad 2011). Consequently, people would be identified accurately through their messages with the ultimate result of reducing online identity theft.

2.3 Educational Data Mining

Data mining is fast becoming a better alternative to surveys in education. Comparatively, it is less expensive to perform than the social science approach of carrying out surveys followed by statistical analysis. 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 environments (Anjewierden et al. 2007). 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 (Romero & Ventura 2010); (Baker & Yacef 2009). Intelligent Multimedia adaptive systems in specific subjects have been proposed with a tutor that can adapt to meet the special needs of the learner (Arroyo et al. 2014); (Walker et al. 2014); (Weragama & Reye 2014); (Westerfield et al. 2015). The third category in this area involves data mining to improve the usefulness of Course Management Systems. Machine learning algorithms have been successfully applied on data from course management systems (CMS) to obtain excellent results as opposed to results obtained from the classroom surveys. Genetic algorithms have also been applied to log data of a web-based learning system to predict students at risk in the early stages (Minaei-Bidgoli et al. 2003) 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 (Minaei-Bidgoli et al. 2003). This can then be used to predict the performance of those students involved.

2.3.1 Sentiment Analysis in Education

² Pavlyshenko, B., Forecasting of Events by Tweet Data Mining. Available at: <http://arxiv.org/pdf/1310.3499>

³ Johansson, F. et al., Detecting Emergent Conflicts through Web Mining and Visualization. Available at: <https://www.recordedfuture.com/assets/Detecting-Emergent-Conflicts-through-Web-Mining-and-Visualization.pdf>

A sentence can either be a fact or an opinion. Opinionated sentences can convey either positive or negative sentiments. Opinions people carry influences 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. Opinion mining will indirectly discover attitudes including emotions of the student towards the course, its lecturer, the environment and colleague students and can be a good indicator of progress academically. Sentiment analysis and opinion mining has been used to determine what people are thinking (Mullen & Malouf 2006) per the issue under discussion. 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 teaching⁴, analysis of student questions and answer board for specific course like Computer Science⁵, Massive Open Online Course (MOOC) Discussion Forums⁶, teacher evaluation policies⁷, and for evaluating Universities (Abdelrazeq *et al.* 2015).

2.4 Performance Prediction

Factors such as accommodation, medium of instruction, parent's educational background, family financial status, place of origin and personal habits are known to affect student's academic performance. Due to this reason, many data mining tasks aimed at predicting academic performance usually ends up with levels of correlation between some attributes and academic performance. Techniques like clustering and classification can be used to place students with likely similar performance in the same cluster⁸. Cluster centers are usually PASS and FAIL. Another approach would be to provide as class labels a division of the students' GPA according to some criteria (Minaei-Bidgoli *et al.* 2003) and then perform classification to determine which student's performance falls within which label. Other researchers try to incorporate personal and pre-university characteristics into the process of predicting academic performance (Kabakchieva 2013). Factors such as the profile of the Senior High School attended, their final score, and score obtained during admission exams are features in the prediction model.

III. Data Collection and Description

The mobile application called WhatsApp is currently becoming the preferred choice of communication (Jisha & Jebakumar 2014) for University students in most developing countries as compared to Twitter and Facebook due to cost and simplicity of use. The application allows users to send secure private messages to their contacts that are also on WhatsApp at a cheaper cost compared to other messaging platforms. To enable many people to have access to messages on the platform, a group of contacts can be created allowing a message to be available to all members of the group. 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, WhatsApp groups exist for teacher unions, student groups, market women, and can be found anywhere 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. Since the application is so popular with students, we decided to collect WhatsApp chats from undergraduate and diploma University students offering Computer Science, Information Technology, Computing with Accounting, and Diploma in Computer Science.

Table 1 Number of messages per option

Program Option	#Students	#Raw Messages	#Cleaned Messages
Computer Science	76	12,332	12,125
Computing with Accounting L300	112	34,834	34,733
Computing with Accounting L400	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

We used the "Email chat" facility in the application to collect the chats from September 2016 to June 2017 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. We deliberately collected messages from higher level classes because they were at that moment doing their core departmental courses actually related

⁴ Sentiment Analysis in Student Experiences of Learning.

Available at: http://sydney.edu.au/engineering/latte/docs/10edm2010_submission_sentimentUSE.pdf

⁵ Sentiment Analysis of a Student Q&A Board for Computer Science.

Available at: http://www.isi.edu/division3/discourse/cmna9_submission_for_camera_final.pdf

⁶ Sentiment Analysis in MOOC Discussion Forums: What does it tell us?

Available at: <http://www.cs.cmu.edu/~mwen/papers/edm2014-camera-ready.pdf>

⁷ Application of Sentiment and Topic Analysis to Teacher Evaluation Policy in the U.S.

Available at: http://www.educationaldatamining.org/EDM2015/uploads/papers/paper_310.pdf

⁸ Classification via clustering for predicting final marks based on student participation in forums. Available at:

http://educationaldatamining.org/EDM2012/uploads/procs/Short_Papers/edm2012_short_3.pdf

to their options. This is reasonable because, students have always complained about borrowed courses not related to their core programs. Including these lower level classes will predictably produce negative sentiments as a result since at the lower level other departmental courses are undertaken.

WhatsApp users can include audio, video or both in their messages. The scope of this work does not cover analysis of multimedia content, therefore multimedia messages were ignored. The chats were heavily infiltrated with phrases of local dialects and jargons. Fortunately for us, most of the English terms that users replaced with jargons and local dialects were stop words such as “*de*” for “*the*”, “*4*” for “*for*”, “*2*” for “*to*” and so on. Notwithstanding, users also replaced some important words required for our algorithms to produce good results with local jargons. 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. 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 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.”

Students were open in sharing their emotions through 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>> is giving them too much pressure and <<name>> not making school an interesting place, apparently due to the perceived pressure. Messages expressing mixed feelings such as the following message were also common:

“Good, I'm happy he said we should be entrepreneurial but sad he didn't say they gonna work hard to restructure our courses and aim more at practical oriented programs to theoretical approach.”

Due to these concerns, we needed to manually pre-process the data before applying our classification algorithms.

3.1 Data Cleaning

To avoid making faulty assumptions and conclusions as a result of the informal nature of messages in our corpus, we tasked two researchers 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 we wanted the sentiments to be author-centric (Balahur & Steinberger 2009) and not around the perception of the reader (Strapparava & Mihalcea 2008), we requested that the researchers do not impose their own sentiments during labeling. 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. During this process, it was discovered that the use of local dialect jargons and English jargons (popularly known as *broken-English*) was so regular in the messages such that we took a decision to train the algorithms with the most common jargons by creating polarity dictionaries (Mohammad et al. 2013) of such jargons as a complement to the training set. Each individual researcher gave labels to the same set of messages and handed over to the second 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 our classification is binary. We as well got rid of messages that had no bearing on education, most of which were religious. The training set was thoroughly reviewed resulting in the labeling of messages as positive or negative.

3.2 Inter-rater Agreement

Since the task at hand is a single-labeled classification, Cohen Kappa (Cohen 1960) was used to calculate the overall agreement between the labeling of the two researchers. In other words, our classification is such that no single data point can fall into more than one class since all classes must be mutually exclusive. Kappa measures the difference between the observed agreement and the expected agreement in the labeling. Its value lies on a -1 to 1 scale, 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.

Table 2 Concordance matrix

<i>2nd observer</i>	<i>1st observer</i>		
	<i>YES</i>	<i>NO</i>	<i>Total</i>
<i>YES</i>	<i>a</i>	<i>b</i>	<i>g₁</i>
<i>NO</i>	<i>c</i>	<i>d</i>	<i>g₂</i>
<i>Total</i>	<i>f₁</i>	<i>f₂</i>	<i>N</i>

$$kappa = \frac{\frac{a+d}{N} - \frac{f_1g_1+f_2g_2}{N^2}}{1 - \frac{f_1g_1+f_2g_2}{N^2}} \quad (1)$$

The overall percent agreement was 0.94 with the Kappa being 0.85 between the annotators. We emphasize here that a high Kappa does not mean the raters are correct about the classes they assign but just an indication that there is a strong agreement between the labels they assign.

IV. Ensemble Sentiment Classifications

We trained and tested three individual classifiers 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. 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). The option of determining sentiments using Lexicon based analysis was not appealing in our case due to the presence of local language and jargons in the chats. This method will parse the data they are given and produce an output by comparing the words they encounter with a finite lexicon of words from a dictionary of opinionated words. However, by training a classifier on a well labeled training set, it is guaranteed to produce acceptable results on similar data during testing.

4.1 Feature Selection for Classification

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. We preprocessed the data 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, we applied tokenization to split the texts at non letters so that anytime a non-letter is encountered the previous characters will denote a new token, therefore splitting the text into words. Though tokenization on non-letters alone may not produce the best outcome for sentiment polarity detection (i.e. since some words on their own are meaningless), we did not have to worry much because majority of the single-worded local jargons were self-sufficient in expressing polarity and therefore was not going to affect the performance of our classification. For words likely to be affected by the problem, we generated n-grams (n being two) after removing tokens smaller than two and larger than fifty. We maintained a minimum token length of two because words like “no” and “ok” can be good determiners of emotions. For emphasis, students relied on the use of repeated characters in words such as “haaaaappppyy”, “haappy”, “yeeaaaah”, etc in their messages. For such repeated characters, we decided to keep only two of the repeated letters in the word. This was to enable us cover all the various forms of such words in the corpus, reduce the dimensionality of our 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 algorithm. 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. To filter stop-words in local jargons, we created a dictionary of words considered to be stop-words in the local language. Their presence is an obvious source of noise that will produce misleading results. We transformed the case of all the tokens to lower case to allow for uniformity. To create the word vector, we made use of the term frequency-inverse document frequency (TF-IDF) (Kotu 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

$$TF = \frac{n_k}{n} \quad (2)$$

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

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

Finally,

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

V. Classification of Student's Messages

5.1 Naïve Bayes

We trained a Naïve Bayes classifier using carefully labeled data from our dataset. With its strong (naïve) independence assumption and simplicity, it performs very well on text classification tasks (MacCallum & Nigam 1998). 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 (Hofmann & Klinkenberg 2014). The concept of NB classification is strongly hinged on allowing every single feature to contribute in the determination of a class (label) for each input text (Bird et al. 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:

$$P(l|f) = \frac{P(l \cap f)}{P(f)} \quad (5)$$

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:

$$P(l|f) = \frac{P(l) * P(f|l)}{\sum_{l \in L} P(l) * P(f|l)} \quad (6)$$

where L is the set of labels used for the classification task.

To avoid the high influence of zero probabilities, we applied the Laplace correction. 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, we made use of Cross-Validation (X-Validation). It partitions the dataset 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 our 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 on the texts.

5.2 Support Vector Machines (SVM)

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 our 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 (Thorsten 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),

$$(x_i y_i), \dots, (x_m y_m) \text{ where } x \in \mathbb{R}^n, \text{ and } y_i \in \{\text{positive, negative}\}$$

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

$$g(x) = \text{sgn}(f(x)) \quad (7)$$

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

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

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

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

$$K(x_i, x) = (x_i \cdot x) \quad (9)$$

with the decision function becoming

$$f(x) = w \cdot x + b \quad (10)$$

where

$$w = \sum_{i=1}^m y_i \alpha_i x_i \quad (11)$$

Training an SVM involves finding solution to α_i and b by optimizing the problem below:

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

subject to the same constraints on the weights α_i .

We leveraged on the effectiveness of SVM and the Linear separable property of texts to classify each input x_i from our 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, we used a complexity constant of 0.001 (tolerance for misclassification) which we think 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.

5.3 Decision Trees

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 we wanted to measure the gain in purity from parent to children down the hierarchy of the tree, we selected gain ratio as the criterion on which attributes will be selected for splitting. The gain ratio (Δ) is given by (Janert 2011)

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

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 aim 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. We used *percentual* pruning 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 we achieved 0.70 accuracy, 0.74 precision, 0.75 recall with an F-score of 0.75.

5.4 Ensemble Classification

Our experiments with the above three classifiers showed that some of the messages were classified differently by different classifiers. In an attempt to combine the strengths of these classifiers, we decided to combine them in an ensemble which will make use of simple majority voting. We trained the ensemble classifier 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 unlabeled 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 (Rokach 2010)

$$class(x) = \arg \max_{c_i \in dom(y)} (\sum_k g(y_k(x), c_i)) \quad (14)$$

where

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

We also implemented another version of this classification by relying on the confidence or probability scores (Asker & Maclin 1997); (Fung et al. 2006) to build the ensemble. Comparing the performance of the two ensemble classifiers, the later underperformed so we opted to use the former for further analysis in this paper. We made use of Split Validation. This was to enable us perform simple validation such that the model will 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 to predict the polarity of unseen data for further analysis.

5.5 Evaluation Measures for Text Classifiers

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.

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)

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

$$Precision(P) = \frac{t_p}{t_p + f_p} \quad (16)$$

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

$$Recall(R) = \frac{t_p}{t_p + f_n} \quad (17)$$

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

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

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.

$$F - score = 2 * \frac{P * R}{P + R} \quad (19)$$

Table 4 shows the comparison of the performance parameters of the classification algorithms used.

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

5.6 Results and Discussions

Using the supervised ensemble classifier, we classified unseen data for each option in Table 1. Fig. 2 depicts the output (polarity of sentiments) produced by the trained ensemble classifier. For instance, the overall sentiment polarity for *BSc. Computing with Accounting level 300* is negative as opposed to *BSc. Computing with Accounting level 400* which is positive. Significant is the fact that, sentiments for all final year classes were positive, whiles the only lower level class produced the only negative sentiments. This could be a fallout from the fact that lower level classes sometimes do “borrowed” courses from other departments as mentioned earlier. We also took notice of distributions of words (activities) that could serve as indicators to students’ emotions in relation to their studies. For example, *student registration* (“*regist_student*”) used to bring a lot of frustration to students due to problems

such as long queues and the inability to register due to non-payment of full fees. In Fig. 3 (a), we observe that the negative emotions were much more than positive emotions towards the process. Other activities such as *examinations* (“*examin*”) and *learning* (“*learn*”) were associated with negative sentiments. The *end of a lecture* (“*lectur_end*”), 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 all the groups. Table 5 shows the list of words, names and phrases serving as direct basis to students’ emotions and sentiments towards the learning process. Worth noting are “*sleep*” and “*assignment*”. In other words, students loaded with several assignments have sleep problems leading to the outpouring of negative sentiments.

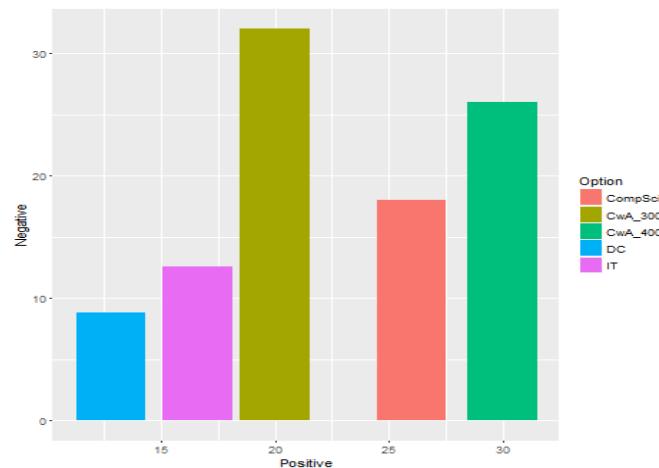


Fig. 2 Sentiment Histograms for the Program Options; for instance, BSc. Computing with Accounting level 300 is negative as opposed to BSc. Computing with Accounting level 400 which is positive

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, Visual Basic and Computer Hardware Architecture.

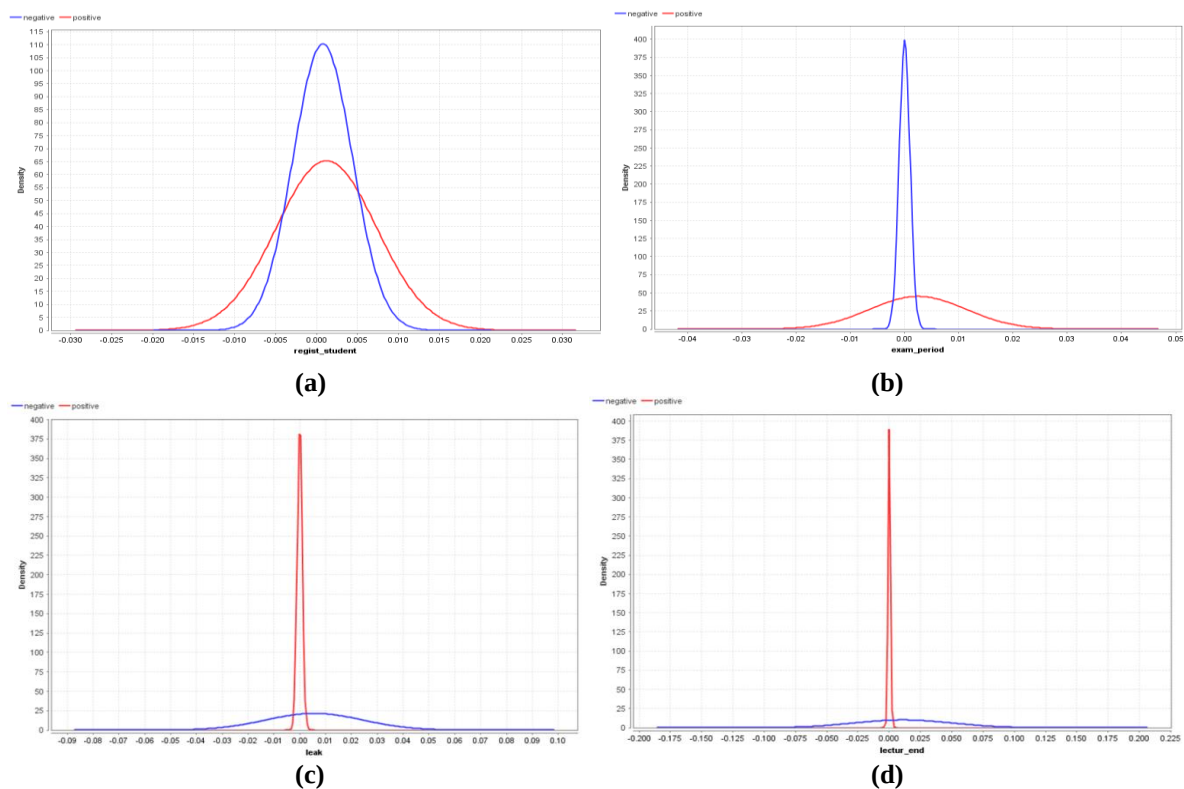


Fig. 3 Probability density functions (pdf) for some words and activities

To obtain a fair idea about the distribution of sentiments over the entire academic year, we classified the sentiments for each program option per month over the said period as depicted in Fig. 4. Our expectation was that, sentiments

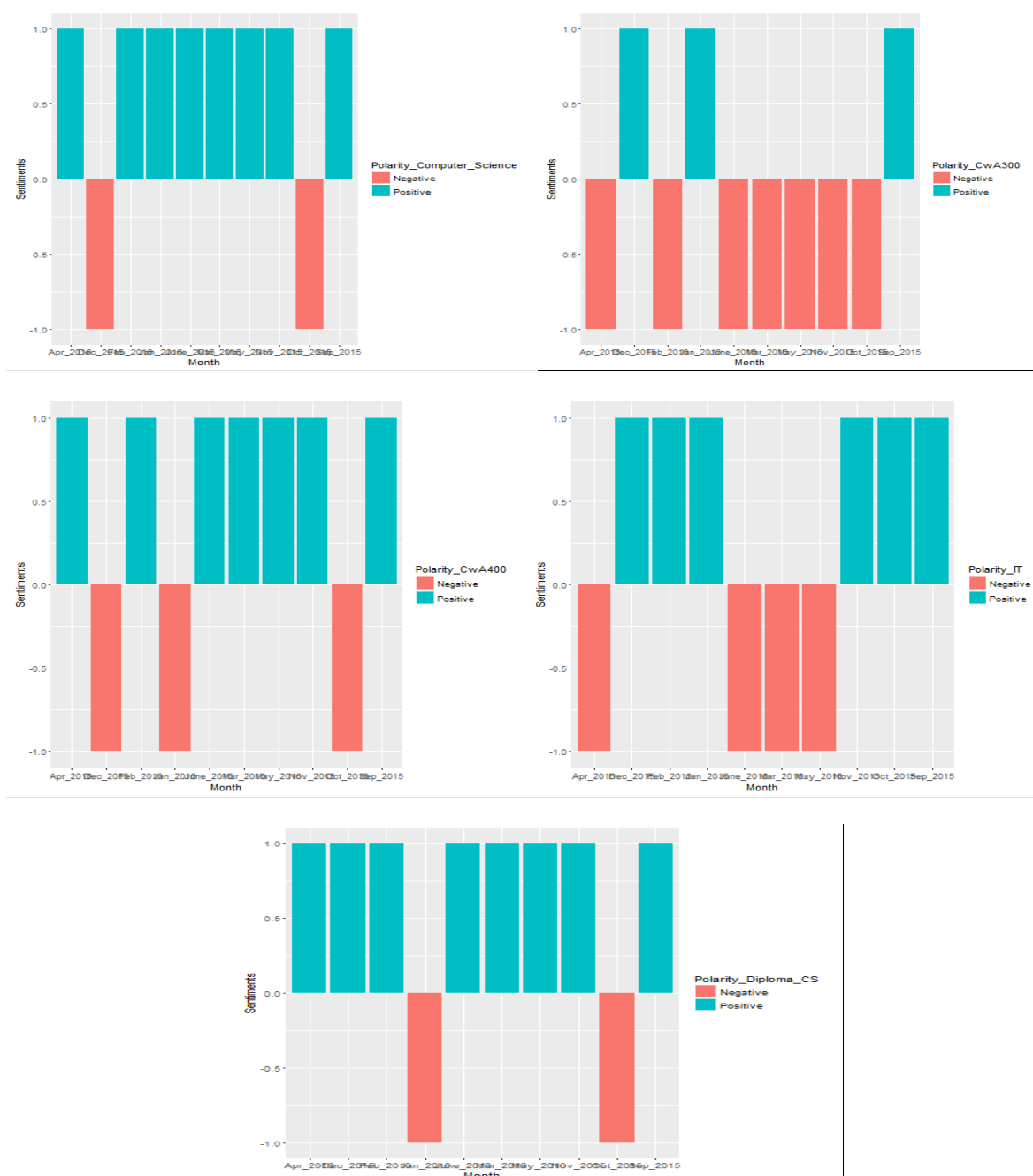


Fig. 4 Monthly Sentiment histograms for the various options. For simplicity, sentiments are considered binary and can be positive (1) or negative (-1) for a given month.

were going to be negative in April since temperatures can go up to 40°C during that period, making studies difficult for the major part of the trimester. However, it turned out that sentiments were not consistently negative for all the options in April, suggesting that either students found a remedy, say in the library, or they didn't border to read much during April.

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

VI. Relationship between Student's Sentiments and Academic Performance

In addition to the classification carried out for each of the program options as a group, we went further to classify the polarity of sentiments for every individual student who contributed by sending a group message. Our objective was to use the academic results and sentiments polarity of each participating student in the group chats to determine via association rule mining if there exist any relationship between a given student's sentiments and their academic performance.

6.1 Association Rule Mining

We wanted to determine if we can identify relationships between student attributes such as *gender*, *program option*, *polarity of sentiments* they express and their *academic performance* based on the outcome of the ensemble classifier and data from students' results database. In association rule mining (also called Market Basket analysis), a relationship (rule) such as

$$\{Gender=Male, Polarity=positive\} ==> Performance=average$$

is interpreted based on the history of occurrences to mean that if a student is Male with positive sentiments, then s/he is likely to be an average student. The strength of this rule can be determined with the use of *support* and *confidence*. Support of an item (e.g. Gender) is the ratio of the frequency of occurrence of the item to the total number of items in the set under consideration. Confidence on the other hand is the accuracy of the association and is given by

$$Confidence (Polarity ==> Performance) = \frac{Support(Polarity \cup Performance)}{Support (Polarity)} \quad (20)$$

Item sets *Gender* and *Polarity* forms the antecedent of the rule with *Performance* being the consequence. A rule is frequent if its support threshold is greater or equal to a specified minimum support. A minimum support is required because as the number of items n increases, the number of item sets increases exponentially by $2^n - 1$ (Kotu 2015), and the number of rules by $3^n - 2^n + 1 + 1$ (Tan et al. 2005). This leads to the generation of many uninteresting rules. *Lift* and *Conviction* are therefore used to filter interesting rules from the possible $3^n - 2^n + 1 + 1$ set of rules.

$$Lift (Polarity ==> Performance) = \frac{Support(Polarity \cup Performance)}{Support (Polarity) * Support (Performance)} \quad (21)$$

If the *lift* of a rule is greater than 1, then the rule is interesting and the antecedent and consequence are positively correlated. A lift less than 1 represents negative correlation between antecedent and consequence and serves as a representation of uninterestingness of the rule. On the other hand, a lift value of 1 shows independence between antecedent and consequence with the rule being uninteresting. *Conviction* will find the ratio of the expected frequency of the antecedent occurring in spite of the consequence in addition to the observed frequency of incorrect predictions.

$$Conviction (Polarity ==> Performance) = \frac{1 - Support(Performance)}{1 - Confidence(Polarity ==> Performance)} \quad (22)$$

We implemented Association Rule mining using the *Apriori* algorithm in WEKA. The *Apriori* method is based on the principle that if an item set is frequent, then all its subsets will also be frequent (Tan et al. 2005). Implementation of *Apriori* involves two main steps. First is the generation of the frequent item sets. This process is computationally intensive than the second process which involves the generation of the rules that are interesting. Since the aim was to find if some form of relationship exists between the polarities of sentiments expressed by students and their academic performance, we retrieved the Grade Point Averages (GPA) of student WhatsApp group members from the University's 2016/2017 first and second trimester result database. Table 5 shows how the GPAs were categorized before applying the *Apriori* algorithm. The algorithm was run with upper bound minimum support of 1 (i.e. 100%) and a lower bound minimum support of 0.1. It starts at the upper bound minimum support and decreases the support by 0.05 (i.e. 5%) until the lower bound min. support is reached or the number of rules required is obtained. Even at 0.1 support, each rule is supported at least by a ceiling of $(0.1 * 211 = 21 \text{ cases})$, which is acceptable for a population of 211 items. A minimum confidence of 0.5 was used since we wanted the algorithm to produce only interesting rules.

Table 6 GPA (x) Categories

Category	GPA Range
Excellent	$4.5 \geq x \leq 5.0$
Good	$3.5 \geq x > 4.5$
Average	$2.5 \geq x < 3.5$
Bad	$x < 2.5$

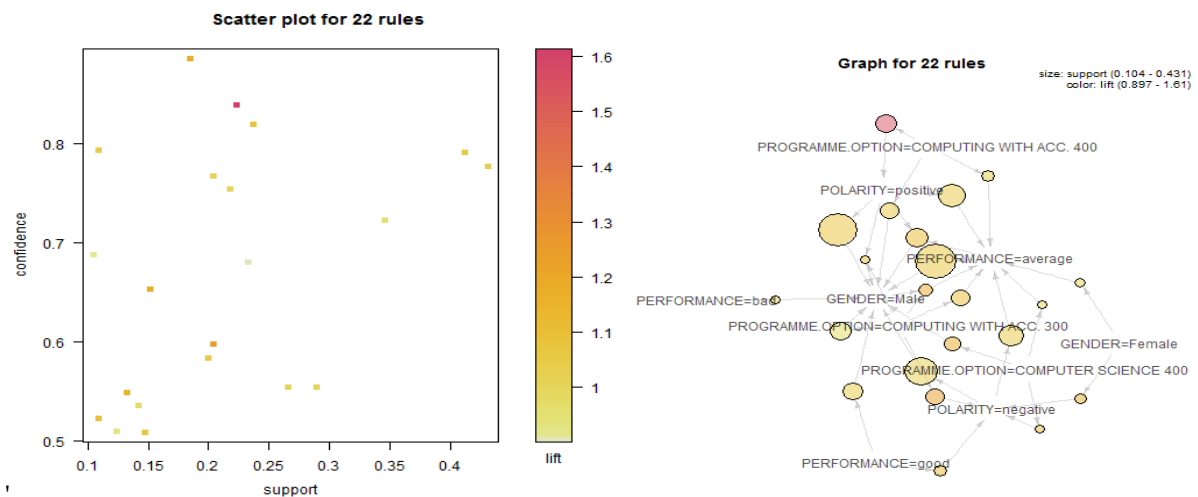


Fig. 5 Scatter Plot and Graph of Non-Redundant Rules

The *Apriori* algorithm generated 45 rules. After removing redundant rules, 22 rules were found interesting as shown in Figure 5. We reproduce five of these rules in Table 7. In rule 3, *confidence* of 0.83 means that 83% of Computing with Accounting 400 students with GPA between 2.4 and 3.5 are likely to exhibit positive sentiments in their messages. A *lift* greater than 1 (in this case 1.6) indicates positive correlation between this Program Option (*Computing with Accounting 400*), its Performance counterpart (*average*) and the consequence sentiment Polarity (*positive*). A 2.39 conviction indicates that the rule would be incorrect 39% more often if the relationship in the rule is based on some random phenomenon. Notwithstanding, this is an instance of an interesting rule which can serve as a basis upon which major educational policy decisions can be taken.

Table 7 Sample association rules produced by the *Apriori* algorithm

Rule #	Antecedent	Consequence	Confidence	Lift	Conviction
3	{Programme Option=Computing with Acc. 400} {Performance = average}	{Polarity = positive}	0.83	1.6	2.39
4	{Polarity = Positive} {Performance = Average}	{Gender = male}	0.82	1.08	1.23
7	{Gender=Male} {Programme Option=Computing With Acc. 400}	{Polarity=Positive}	0.79	1.52	2.06
11	{Polarity=Positive} {Performance=Good}	{ Gender=Male}	0.77	1.01	0.91
27	{Gender=Male} {Polarity=Positive}	Performance=Average}	0.57	1.04	1.02

VII. Conclusions and Future Work

In this paper, we have successfully used supervised ensemble classification to predict the polarities of sentiments expressed by University students towards their course work in their WhatsApp group messages. An F-score of 0.80 was obtained for the classification, the output was then used successfully to determine the various interesting relationships existing between the Gender of the student, their program option, polarity of sentiments they express and their academic performance. This will serve as a guiding tool to educational policy makers to make informed decisions that will incorporate students' emotions into decision making.

As future work, we would want to predict the extent to which students' usage of WhatsApp jargons affect their academic writings (performance).

VIII. References:

- Abdelrazeq, A. et al., 2015. Sentiment Analysis of Social Media for Evaluating Universities. In *Proceedings of Second International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2015)*, Dubai, UAE. pp. 49–62.
- Ali, T. et al., 2013. Can I hear you? Sentiment Analysis on Medical Forums. In *International Joint Conference on Natural Language Processing*. pp. 667–673.
- Anjewierden, A., Bas Kolloffel & Hulshof, C., 2007. Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. In *Proceedings of the International Workshop on Applying Data Mining in e-Learning*. pp. 23–32.
- Arroyo, I. et al., 2014. A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. *Int J Artif Intell Educ*, 24, pp.387–426.

- Ashktorab, Z., Brown, M. C. and Nandi & A., C., 2014. Tweedr: Mining Twitter to Inform Disaster Response. In *Proceedings of the 11th International ISCRAM Conference—University Park, Pennsylvania, USA*.
- Asker, L. & Maclin, R., 1997. Ensembles as a sequence of classifiers. In *Proc. of IJCAI*. pp. 860–865.
- Baker, R. & Yacef, K., 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. *J. Educational Data Mining*, 1(1), pp.3–17.
- Balahur, A. & Steinberger, R., 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. In *Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis*.
- Bird, S., Klein, E. & Loper, E., 2009. *Natural Language Processing with Python*, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopo.
- Boom, N. ten, 2015. Predicting Short Term Company Performance by Applying Sentiment Analysis and Machine Learning Algorithms on Social Media. In *23th Twente Student Conference on IT*. pp. 1–6.
- Carchiolo, V., Longheu, A. & Malgeri, M., 2015. Using Twitter Data and Sentiment Analysis to Study Diseases Dynamics. *Springer International Publishing Switzerland*, pp.16–24.
- Chen, X., Vorvoreanu, M. & Madhavan, K., 2014. Mining Social Media Data for Understanding Students' Learning Experiences. *IEEE Transactions on Learning Technologies*, 7(3), pp.246–259.
- Chikersal, P. et al., 2015. Modelling Public Sentiment in Twitter: Using Linguistic Patterns to Enhance Supervised Learning. *Springer International Publishing Switzerland*, pp.49–65.
- Choudhury, M.D. et al., 2013. Predicting Depression via Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence*. pp. 128–137.
- Cohen, J., 1960. A Coefficient of Agreement for nominal scales. *Educational and Psychological Measurement*, 20, pp.37–46.
- Deng, Y., Stoehr, M. & Denecke, K., 2014. Retrieving Attitudes: Sentiment Analysis from Clinical Narratives. *MedIR*, pp.12–15.
- Fung, G.P.C. et al., 2006. A balanced ensemble approach to weighting classifiers for text classification. In *Proc. of ICDM*. pp. 869–873.
- Gamallo, P. & Garcia, M., 2014. Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. pp. 171–175.
- Gilbert, E. & Karahalios, K., 2009. Predicting Tie Strength With Social Media. *CHI*.
- Goffman, E., 1959. *The Presentation of Self in Everyday Life.*, Lightning Source Inc.
- Harer, S. & Kadam, S., 2014. Mining and Summarizing Movie Reviews in Mobile Environment. *International Journal of Computer Science and Information Technologies*, 5(3), pp.3912–3916.
- Hofmann, M. & Klinkenberg, R., 2014. *Data Mining Use Cases and Business Analytics Applications* V. Kumar, ed., Taylor & Francis Group, CRC Press.
- Janert, P.K., 2011. *Data Analysis with Open Source Tools*, O'Reilly Media, Inc. 1005 Gravenstein Highway North, Sebastopol.
- Jisha, K. & Jebakumar, 2014. Whatsapp: A Trend Setter in Mobile Communication among Chennai Youth. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS)*, 19(9), pp.1–6.
- Kabakchieva, D., 2013. Predicting Student Performance by Using Data Mining Methods for Classification. *Bulgarian Academy Of Sciences, Cybernetics And Information Technologies*, 13(1), pp.61–72.
- Kiritchenko, S., Zhu, X. & Mohammad, S.M., 2014. Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, 50, pp.723–762.
- Kotu, V., 2015. *Predictive Analytics and Data Mining Concepts and Practice with RapidMiner* S. Elliot, ed., Elsevier Inc.
- MacCallum, A. & Nigam, K., 1998. A comparison of event models for naïve bayes text classification. In *AAAI-98 workshop on learning for text categorization*.
- Matykievicz, P., Duch, W. & Pestian, J.P., 2009. Clustering semantic spaces of suicide notes and newsgroups articles. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, BioNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA.*, pp. 179–184.
- Mensah, P.K., 2015. Textual prediction of attitudes towards mental health. *Int. J. Knowledge Engineering and Data Mining*, 3(3/4), pp.274–285.
- Mensah, P.K. & Akobre, S., 2015. Mining Social Media for Conflict Prevention and Resolution. *African Journal of Computing & ICT*, 8(3), pp.163–170.
- Merceron, A. & Yacef, K., 2003. A Web-Based Tutoring Tool with Mining Facilities to Improve Learning and Teaching. *Artificial Intelligence in Education, H.U. Hoppe et al. (Eds.)*, IOS Press, pp.201–208.
- Minaei-Bidgoli, B. et al., 2003. Predicting Student Performance: An Application of Data Mining Methods With An Educational Web-Based System. In *33rd ASEE/IEEE Frontiers in Education Conference*. pp. 13–18.
- Mohammad, S.M., 2011. *Sentiment Analysis of Mail and Books*, National Research Council Canada.
- Mohammad, S.M., S. Kiritchenko & Zhu, X., 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proc. of SemEval 2013*. pp. 321–327.
- Moreira, S. et al., REACTION: A naïve machine learning approach for sentiment classification. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Seventh International Workshop on Semantic*.
- Mullen, T. & Malouf, R., 2006. A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI/CAAW)*. pp. 159–162.
- Pestian, J.P., Matykievicz, P. & Grupp-Phelan, J., 2008. Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, BioNLP '08, Association for Computational Linguistics, Stroudsburg, PA, USA*. pp. 96–97.
- Porter, 1980. An algorithm for suffix stripping. *Program*, 14(3), pp.130–137.
- Rokach, L., 2010. Ensemble-based classifiers. *Artif Intell Rev*, 33, pp.1–39.
- Romero, C. & Ventura, S., 2010. Educational Data Mining: A Review of the State of the Art. *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Rev.*, 40(6), pp.601–618.
- Roosevelt, M., 2012. Social Media Analytics: Data Mining Applied to Insurance Twitter Posts. *Casualty Actuarial Society E-Forum*, 2, pp.1–36.
- Sassano, M., 2003. Virtual Examples for Text Classification with Support Vector Machines. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. pp. 208–215.
- Sokolova, M. & Bobicev, V., 2011. Sentiments and Opinions in Health-related Web Messages. In *Proceedings of Recent Advances in Natural Language Processing*. pp. 132–139.
- Strapparava, C. & Mihalcea, R., 2008. Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM symposium on Applied Computing*.
- Tan, P.-N., Steinbach, M. & V, K., 2005. *Association Analysis: Basic Concepts and Algorithms.*, Addison Wesley.

- Thorsten, J., 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning, Springer Verlag, Heidelberg, DE*. pp. 137–142.
- Walker, E., Rummel, N. & Koedinger, K.R., 2014. Adaptive Intelligent Support to Improve Peer Tutoring in Algebra. *Int J Artif Intell Educ*, 24, pp.33–61.
- Wan, Y. & Gao, Q., 2015. An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. In *IEEE 15th International Conference on Data Mining Workshops*. pp. 1318–1325.
- Weragama, D. & Reye, J., 2014. Analysing Student Programs in the PHP Intelligent Tutoring System. *Int J Artif Intell Educ*, 24, pp.162–188.
- Westerfield, G., Mitrovic, A. & Billinghamurst, M., 2015. Intelligent Augmented Reality Training for Motherboard Assembly. *Int J Artif Intell Educ*, 25, pp.157–172.
- Yano, T., Smith, N.A. & Wilkerson, J.D., 2012. Textual Predictors of Bill Survival in Congressional Committees. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 793–802.
- Yeboah, J. & Ewur, G.D., 2014. The Impact of Whatsapp Messenger Usage on Students Performance in Tertiary Institutions in Ghana. *Journal of Education and Practice*, 5(6), pp.157–164.