RESEARCH ARTICLE OPEN ACCESS

Unilife: A Mobile Application for Assisting University Students to Cope With Study and Social Activities Using Machine Learning Algorithms

Shatha Alsaedi
School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane
Australia

ABSTRACT
With the ongoing rapid development of mobile phone technology and use, mobile phone applications have become more and more popular, and their usage has increased exponentially. In order to achieve a balance between students' smartphone usage and study time, we implemented the UniLife smartphone application, which efficiently connects students with their tutors for better educational outcomes. Students post their needs directly in the app*lication*'s forum, and as the relevant digital data increase in volume, there is an urgent need to apply machine learning algorithms to analyze these data. VADER Sentiment Analysis and LDA topic modeling algorithms have been used in this application to analyze students' posts, report sentiments, and model the discussed topics to tutors so that they become aware of their students' needs and issues. As a consequence, the tutors can overcome these problems, enhance the educational process, and help students achieve better study results.
Keywords :— LDA, sentiment analysis, learner opinion, posts, VADER.

I. INTRODUCTION

With the ongoing rapid development in mobile technology, mobile phone applications have become more popular. Their usage has increased exponentially. Nowadays, a significant number of people use mobile applications, where over 170 million apps were downloaded in 2015 [6]. Students comprise an important segment of applications' users. In order to help students to achieve higher results in their study, it is useful to connect their study process with the usage of mobile applications. The goal of this research is to enhance the educational process by conveying students' feedback, about the courses, to their tutors.

In recent years, students have been using website-based forums to convey their feedback, problems, and suggestions concerning their study courses. However, there has been a considerable shift from website uses to mobile apps uses. This is caused by the fact that apps uses are generally easier and more convenient. Accordingly, some students show reluctance to use websites and tend to replace them with mobile apps. Hence, further significant efforts should be devoted to develop mobile apps. This study aims at creating a mobile application forum to help students engage in their study through writing study-related posts, and report their contents to their tutors. In fact, such posts are beneficial for the educational process, but the amount of data has been significantly increasing rendering it very difficult to be analysed manually. Data mining should be applied to study such data [1]. However, the nature of such posts may comprise a challenge to the analysis task, because the posts usually contain slangs and emoji characters [5].

In this research, an interactive iOS application is implemented, namely UniLife, which contains students' discussions forum. Moreover, topic modelling, and sentiment analysis algorithms, are applied to extract hidden topics within posts, and analyse students' sentiments, from students' discussion, and report them to their tutors.

II. BACKGROUND

A. Natural Language Processing (NLP):

The use of computer technology, to understand and analyse human language, started circa 40 years ago, where it was called natural language understanding. In fact, it was later called natural language processing (NLP), which is a branch of artificial intelligence (AI). However, NLB has been taught in computational statistics and mining text data. It is mainly used to analyse texts, as well as speech, and it is usually used to discover semantic similarity and language translations [15].

The most important processes in NLP is understanding the text semantics. This is associated with many difficulties such as word ambiguity and various grammars. To achieve this goal, there are many approaches to determine word meaning, tagging, and parsing.

B. Text analytics:

Data posted in many websites, and microblog platforms, have been increasing dramatically in a way that makes it very difficult to analyse them manually, and the results of such a process are very beneficial for the society and other useful

applications, and it becomes very important to find a way to analyse the data automatically using text analytics algorithms [5].

Text analytics are mechanisms that apply machine learning and computational procedures on texts to extract useful data from the text, and to use them in fields of research, education, commerce, and other useful applications [3].

C. Sentiment Analysis:

All texts can be categorised into two main parts of facts (objective) and opinions (subjective) [16]. The facts are objective about some entities, while the opinions are subjective describing how people feel or think (their sentiments) towards those entities. Given the fact that digital data have been rapidly increasing recently, relevant research has focused on sentiment analysis, and finding a mechanism to analyse it computationally.

1) Sentiment analysis definition:

Sentiment analysis is a process of computationally extracting opinions and sentiments from the text stream [17]. It is very important to take into consideration the sentiment analysis because it carries important information, and is not an easy process even for humans because extracting opinions by deciding whether the words are positive or negative is not enough, and it should be known what entities such words describe to appropriately determine the relevant opinion. For example, the word "fast" in the sentence of "This CPU is fast" is of positive nature, while in the sentence "This battery drains fast" is of negative nature [2].

D. Topic modelling:

Recently, as the textual information being digitalized while associated with massive amount of data, it becomes very difficult for humans to manually find what they are looking for within such data, or to identify what all this information is related to [8]. Therefore, there is an urgent need to analyse the text automatically, particularly when documents become more complex [21].

1) Topic modelling definition:

The topic discovery is a statistical model that uses data mining techniques and algorithms to extract topics embedded in the text. Technically, a topic is a probabilistic distribution over the text, in other words, the topic is a semantic theme of the document [9].

In fact, topic model algorithms usually focus on discovering frequency of words appearance together within the text, and the topics can be detected by computing the highest probabilities of the related words [12]. Some documents can represent one specific topic, while others can have a combination of many topics.

Topic modelling is not exclusively used in text data, rather it can be used in many other applications such as computer vision [12].

III. RELATEDWORK

Given the fact that users' posts in social media websites, and finding interests of the users, or analyzing those posts to discover the latent topics inside them, are issues of research previously carried out, and there are many existing studies that discussed such issues, where each of them has its unique way of finding sentiments, or modeling topics, or both at the same time.

The next section highlights some existing research papers in this domain, and for each one the key features of it's algorithm is briefly illustrated. First, we will present some existing works in topic discovery, and second, we will highlight some research papers on opinion mining. Finally, we will present some studies on algorithms for text analysis, and topic and sentiment generation at the same time.

A. Related work in topics discovery:

There are considerable research efforts to model the topics using either supervised or unsupervised methods. Some of them interested in extracting topics from long, well structured and grammatically correct text while other concerned about short, noisy and unstructured text. In fact, our project interested in later one because this is the nature of the students' posts in the discussion forums. Related to existing topic modeling research, the next section reviews work which focuses on topic modelling in microblogs social media and learning forums data.

1)Topics discovery in microblogs social media:

Vicient and Moreno [10] focus in their study on extracting topics from microblogs nature text using unsupervised method, unlike some other research works which are basically syntactic. Their work traits microblogs analysis in different way because the text in these websites is different from other text as it is unstructured, has some spelling and grammar mistakes, noisy and short. They proposed a novel topic modelling algorithm which involves mapping of hashtags to WordNet terms and their posterior clustering with respect to the semantic. Their analysis process contains three main steps which are: semantic annotation, semantic clustering and topic selection.

In addition to social media websites, the topics can be extracted from any source. Although there are many differences between short and long text which required different processing, the researchers focus on both types of text but the need to analysis short text has increased exponentially because nowadays, most text has social media websites nature.

Moreover, Chen and Liu [7] have proposed a new algorithm which can be used in topic discovery. Unlike other topic discovery algorithms which need wide documents to detect the topics, this algorithm can work in small documents as well. This approach which is used in the algorithm works correctly by using previous knowledge. It generates two form of knowledge must-link (meaning that two words should be in the same topic) and cannot-link (meaning that two words should not be in the same topic).

2)Topics discovery in learning forums:

analyzing learners' posts in learning forms has high priority as it generates useful outcomes for learners, educators and educational process in general. These forums can be analyzed using many types of algorithms depending on researcher's perspective and domain needs. There are considerable works on this issue some of them interested in finding latent topics inside learners' discussions, while others focus on analyzing their sentiment, using these finding in recommendation system or applying different algorithms to achieve all of these goals in related fashion. The next section shows some of existing works in this domain which focus on applying machine learning algorithms on learning environment posts.

Luo *et al.* [4] research interests in mining the learning posts because they motivated by the fact that learner's post directly highlights his need and they analyzed these posts to extract the topics and searching for learners who have same interests and recommend for them some learning resources using recommendation system. after finding latent topics inside learners' posts, they use these results in recommendation system, they use a recommendation algorithm to recommends the learners with the same interest and suitable learning resource.

In addition, Daniil, Dascalu, and Trausan-Matu [1] focus on applying machine learning algorithms on Computer Supported Collaborative Learning (CSCL) environment. They proposed a novel algorithm to automatically analyze posts and discussion threads in learning forum as well. Their approach uses a multi-layered architecture. What makes this architecture effective that the algorithm groups into levels, in each one there are key operation related to this layer. Moreover, each level is traded independently and if any error occurs it can easily be diagnosed and solved. Furthermore, in case of unsatisfied results, it can be easily go back to big error location and corrected.

B. *Related work in opinion mining:*

Analyze text to detect opinion of it's writer as a research problem which has been existing previously. The need of this kind of analysis has many important applications. Researchers focus in extracting the sentiment by different methodologies and algorithms each of them has its own unique strengths. The next section present some of them briefly.

Jiang *et al.* [11] have interests in social media text analysis, they proposed a novel algorithm, namely emoticon space model (ESM), which projects words and social media posts into an emoticon space in order to specifies the subjectivity, polarity and emotion. In addition, there is an existing work on opinion mining using multinomial Na¨ive Bayes classifier. Pak and Paroubek [14] work on an algorithm which automatically collects the corpus and using these corpora the classifier is built and it has the ability to determine positive, negative and neutral opinions in the document.

C. *Related work in topic modeling and opinion mining at the same time:*

Some researchers believe in the importance of finding algorithms which detect the hidden topics and analyze the subjectivity of text at the same time. Thus, there are existing works which mainly focus in this issue. The next section presents some of theses works.

Lin and He [3] and Lin *et al.* [2] overcome some problems of previous work on sentiment analysis such as the domain dependency in trained classifiers, the labeled corpora which is difficult to apply in real world text and the problem in which these algorithms do not take into consideration the relation ship between sentiment and topic. The authors have proposed new family of probabilistic topic models, namely joint sentiment/topic models (JST), which can detect sentiment in connection with topic without requiring text labeled.

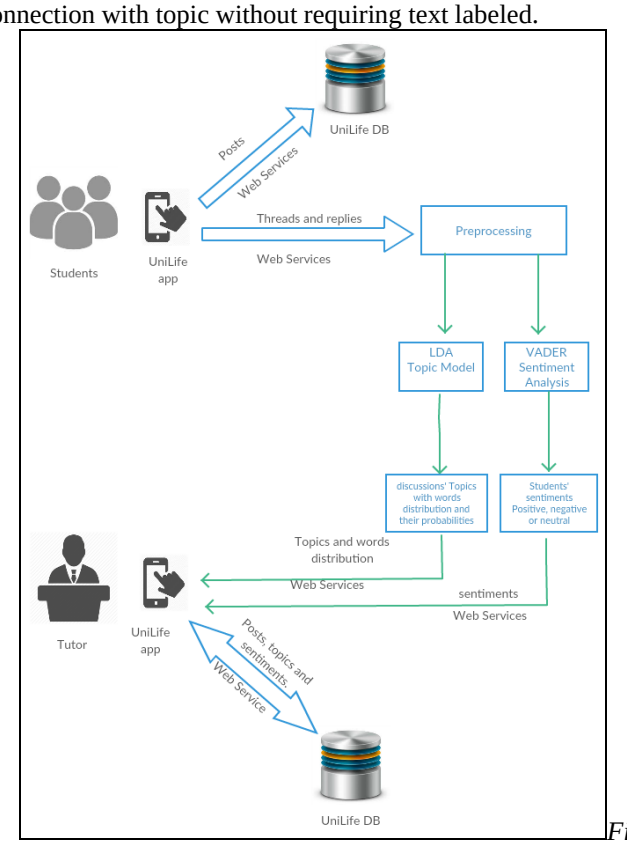

Fig 1. Proposed framework architecture.

IV. METHODOLOGY

Our project consists of two main parts: front end and back end. Front end implementation focuses on the application features and how it deals with server side, web services and remote database. Firstly, in front end, the iOS application is developed including many features which is: 1. Students and tutors sign up using their new account in the application, their Facebook account, or their Twitter account. 2. After creating a valid account, students can post their needs or problems directly into the application's forum or can browse other students' threads and replies. A web service manages the data

transfer from the application to the remote database and the server, which is responsible for performing the data analysis.

Secondly, all posts are stored in remote database through web service to be able to be retrieved at any time if the user browses the forum's contents.

Thirdly, all posts are transferred to the server which are responsible in applying machine learning algorithms on posts. In fact, there are two web servers one of them are responsible in applying sentiment analyses algorithm wilt the other performs topic model algorithm.

In the first web server, posts are pre-processed to prepare the text for analysis step, then VADER sentiment analyses algorithm is applied on all replies. In addition, the result which is student's sentiment is transferred to the remote database to be saved and able to viewed in the forum.

In addition, in the second web server, posts are preprocessed and LDA topics model is applied on them to extract the hidden themes in students' discussions. The discovered topics is associated with words distributions along with their probabilities. These information is stored in the database as well. Then it, along with students' sentiment, are transferred in JSON format through web service to be viewed in the smart phone application by the tutor.

Moreover, the application contains other features such as local search engine to search the universal database using keywords. Figure. 1 illustrates the architecture of UniLife application.

VADER sentiment analysis algorithm:

VADER (Valence Aware Dictionary and sEntiment Reasoner) [19] is lexicon rule based algorithm which is proposed to analyse the sentiment of microblog media like social media content and this is one of the main reasons of considering this algorithm to analyse students' sentiments in our smartphone app. In fact, Vader sentiment lexicon works better in microblogging platform text nature because it contains lexical items for some words which carry sentiment like emoticons, slangs, initialisms and acronyms which has been used widely in social media text as well as it accounts the differences of the intensity of sentiment for words [19].

Eventually, after applying this algorithm on students' post, each sentence result compound sentiment intensity polarity which range from +4 for very positive and -4 for very negative and 0 for neutral.

latent Dirichlet allocation (LDA):

LDA [18] is a probabilistic topic model algorithm which extracts the hidden themes in the document. It represents each document as a mixture of topics each one is descried by number of top words. It results a matrix of distribution of topics in documents and distribution of words in a topic. It is unsupervised algorithms and it does not need to training dataset.

V. EXPERIMENT

Dataset: A students' forum posts are used to test the application's performance. Threads and replies are collected using specific code which is written in Python programming language, the input is the forum's URL and the output are files in JSON format each of them represents one thread along with all it's replies and other related information. There was a need to proposed a special algorithm, and implement it using Python language, which be able to read all these files, extract threads contents, connects them with their replies and upload them to the remote database to be viewed in the application's forum using web services.

Fig 2. Topics interface.

Pre-processing: all posts are pre-processed before they can be analysed, different pre-processing steps are applied on posts based on whether they are analysed to modelling topics or sentiment detection. In more details, in sentiment analysis for example, opinion detection is influenced by booster words or emojis while in topis extraction these do not effect topics modelling and some of them removed during pre-processing step.

Framework development: The iOS application is implemented using Xcode IDE and the database is implemented using SQL server. In addition, web services are programmed in C# using Microsoft Visual Studio. Both LDA topic model and VADER sentiment algorithms implementations are programmed in Python as open source codes.

As the forum is accessed by all users no matter where they are, the database should be accessed from any where and Smarterasp.com is used to host the database to be accessed using web service. Moreover, once student post a thread or replay, it should be dynamically analyzed to result hidden topics or student's opinion and this should be processed in server side using cloud computing and processed results are transferred to the database to be accessed by the app.

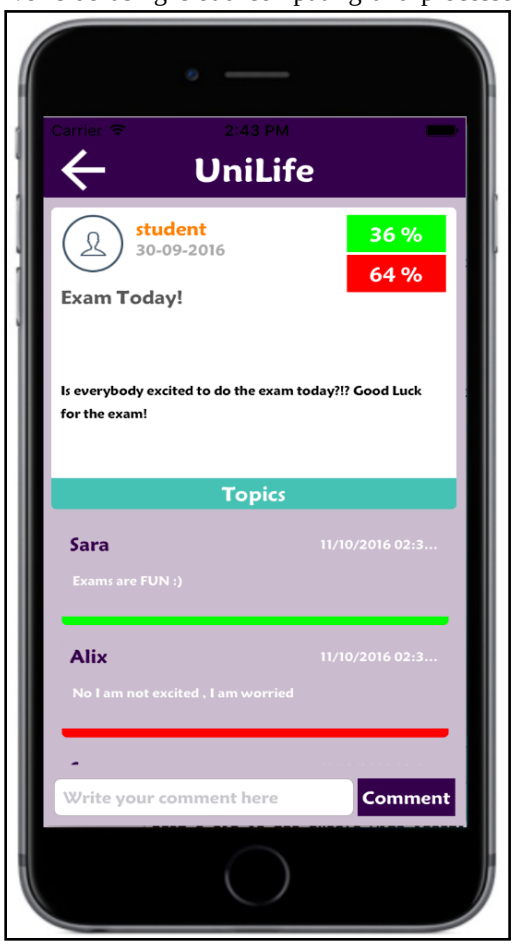

Fig 3. Students' sentiments interface

VI. RESULTS AND OBSERVATION

To testing the performance of our proposed application, about 3000 posts from students' forum are uploaded to the application's database. In fact, as LDA topics, each topic is represented by words distributions along with their probabilities. To facilitate understanding topics for tutors, top words for each topic are presented without viewing their probabilities in the application's interface and these top words appear in order according to their probabilities.

Figure 2 shows some of extracted topics, topic 0 words are (mark, manual, point, effort) and this topic may represent exam marking method. In addition, topic1 words are (point, answer, undergraduate, time, exam) and this topic may represent the total exam time.

In addition, each student's posts are analysed and detected whether it represents positive, neutral or negative opinion and according to this it marked in green, grey or red respectively.

Eventually, the total percent of students' opinion is calculated and presented in the thread interface, Figure 3 illustrates sentiment analysis interface, as presented in this Figure, students are commenting regarding the final exam. By reading the comments, the course's tutor can note the overall negative opinion and address the students' worries and anxieties about the final exam. If he wants more details, he can look through the red labelled comments to find out exactly what the main reasons for their fear are in order to provide suitable help, such as offering additional tutorial sessions or exam exercise materials.

In fact, as some of students' posts are very short, there may be some irregular topic's words appear in some cases. To enhance results more, threads and their replies can be joined together as one document input to LDA model. While LDA works better for long, well structured and formal texts, the results can better if Twitter-LDA model is used because students' posts are relatively similar to tweets in nature, both are short and usually one topic is discussed in each post.

VII. CONCLUSION

In this paper, we have proposed the UniLife iOS application using machine learning algorithms. This application works as a bridging tool by connecting students and their tutors and provides many services for them. It facilitates students' interaction between each other and between them and their tutors. VADER sentiment analysis and LDA topic model algorithms have been applied in the application's forum to explore hidden topics in students' posts, summarize their opinion towards these topics and report them to their tutor. This process helps tutor in finding students' weakness and problems in short time and overcome these problems.

ACKNOWLEDGMENT

The author would like to thank her supervisor, Dr. Helen Huang, for her supervision, guidance and support during the work on this research. Besides, she would like to thank Taibah University for sponsorship her study at The University of Queensland.

REFERENCES

[1] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, and A. Gupte, "Comparative study of classification algorithms used in sentiment analysis," *IJCSIT) International Journal of Computer Science and Information Technologies*, vol. 5, pp. 6261-6264, 2014.

[2] C. Lin, E. Ibeke, A. Wyner and F. Guerin, "Sentiment-topic modeling in text mining", *WIREs Data Mining Knowl Discov*, vol. 5, no. 5, pp. 246-254, 2015.

[3] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceedings of the 18th ACM*

conference on Information and knowledge management, 2009, pp. 375-384.

[4] C. Luo, T. He, X. Zhang, and Z. Zhou, "Learning Forum Posts Topic Discovery and Its Application in Recommendation System."

[5] M. Kumar, "AUTOMATIC IDENTIFICATION OF USER INTEREST FROM SOCIAL MEDIA.", Master of Computer Science, Dalhousie University, Halifax, Canada,2015.

[6] "What is a WebView? -", *Telerik Developer Network*, 2015. [Online]. Available: http://developer.telerik.com/featured/what-is-a-webview/. [Accessed: 04- Apr- 2016]

[7] Z. Chen and B. Liu, "Mining topics in documents: standing on the shoulders of big data," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1116-1125.

[8] D. Blei, "Probabilistic topic models", *Communications of the ACM*, vol. 55, no. 4, p. 77, 2012.

[9] D. Blei and J. McAuliffe. Supervised topic models. In Neural Information Processing Systems, 2007.

[10] C. Vicient and A. Moreno, "Unsupervised topic discovery in micro-blogging networks", *Expert Systems with Applications*, vol. 42, no. 17-18, pp. 6472-6485, 2015.

[11] F. Jiang, Y. Liu, H. Luan, J. Sun, X. Zhu, M. Zhang and S. Ma, "Microblog Sentiment Analysis with Emoticon Space Model", *J. Comput. Sci. Technol.*, vol. 30, no. 5, pp. 1120-1129, 2015.

[12] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categoriesIEEE Computer Society Conference on Computer Vision and Pattern Recognition", in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524-531.

[13] Y. Hu, J. Boyd-Graber, B. Satinoff and A. Smith, "Interactive topic modeling", *Mach Learn*, vol. 95, no. 3, pp. 423-469, 2013.

[14] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", in *LREc*, 2010, pp. 1320-1326.

[15] A. Martinez, "Natural language processing", *WIREs Comp Stat*, vol. 2, no. 3, pp. 352-357, 2010.

[16] B. Liu, "Sentiment Analysis and Subjectivity", *Handbook of Natural Language Processing*, vol. 2, pp. 627-666, 2010.

[17] P. Pantone, "Adding Sentiment Analysis support to the NLTK Python Platform", Master of Science, University of Edinburgh, 2015.

[18] D. Blei, A. Ng and M. Jordan, "Latent dirichlet allocation", *Journal of machine Learning research*, vol. 3, pp. 993-1022., 2003.

[19] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.", in *Eighth International Conference on Weblogs and Social Media*, 2014