ISSN:0975-9646

Survey on Analysis of Students' Emotions through Social Media Data Mining

Ranjeeta Rana[#1], Mrs. V. L. Kolhe[*2]

[#]Department of Computer Engineering, D. Y. Patil College of Engineering,
Akurdi, Pune, Savitribai Phule Pune University, India

[*]Department of Computer Engineering, D. Y. Patil College of Engineering,
Akurdi, Pune, Savitribai Phule Pune University, India

***Abstract*— Students' informal conversations on social media(e.g Twitter, Facebook) shed light into their educational experiences-opinions, feelings, and concerns about the learning process. Data from such un-instrumented environments can provide valuable knowledge to inform student learning. Analyzing such data, however can be challenging. The complexity of students' experiences reflected from social media content requires human interpretation. However, the growing scale of data demands automatic data analysis techniques. Twitter posts of engineering students' is focused to understand issues and problems in their educational experiences. Engineering students encounter problems such as heavy study load, lack of social engagement and sleep deprivation are considered. A multi-label classification algorithms to classify tweets reflecting students' problems is implemented.**

***Keywords*— Education, computers and education, social networking, web text analysis.**

I. INTRODUCTION

Social networks have become very popular in recent years because of the increasing proliferation and affordability of internet enabled devices such as personal computers, mobile devices and other more recent hardware innovations such as internet tablets. In general, a social network is defined as a network of interactions or relationships, where the nodes consist of actors, and the edges consist of the relationships or interactions between these actors. A generalization of the idea of social networks is that of information networks, in which the nodes could comprise either actors or entities, and the edges denote the relationships between them. Social media sites such as Twitter, Facebook, and YouTube provide great venues for students to share joy and struggle vent emotion and stress and seek social support. On various social media sites , students discuss and share their everyday encounters in an informal and casual manner. Students' digital footprints provide vast amount of implicit knowledge and a whole new perspective for educational researchers and practitioners to understand students' experiences outside the controlled classroom environment[1]. This understanding can inform institutional decision-making on interventions for at-risk students, improvement of education quality, and thus enhance student recruitment, retention, and success. The abundance of social media data provides opportunities to understand students' experiences but also raises methodological difficulties in making sense of social media data for educational purposes. Traditionally, educational researchers have been using methods such as surveys, interviews, focus groups, class room activities to collect data related to students learning experiences. These methods are usually very time consuming, thus cannot be duplicated or repeated with high frequency. The scale of such studies is also usually limited. In addition, when prompted about their experiences, students need to reflect on what they were thinking and doing sometime in the past, which may have become obsured over time.

The research goals are 1) to demonstrate a workflow of social media data sense-making for educational purposes, integrating both qualitative analysis and large-scale data mining techniques, 2) to explore engineering students informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences.

Data mining can be defined as the process involved in extracting interesting, interpretable, useful and novel information from data. It has been used for many years by businesses, scientists and governments to sift through volumes of data like airline passenger records, census data and the supermarket scanner data that produces market research reports. The objective of data mining in each application area is different. For example, in business the main objective is to increase profit, which is tangible and can be measured in term of amounts of money, number of customers and customer loyalty[19]. Traditionally educational researchers have been using methods such as surveys, interviews, focus groups, classroom activities to collect data related to students learning experiences. These methods are usually very time consuming, thus cannot be duplicated or repeated with high frequency. The scale of such studies is also usually limited. In addition , when prompted about their experiences, students need to reflect on what they were thinking and doing sometime in the past, which may have become obsured over time. The emerging field of learning analytics and educational data mining has focused on analyzing structured data obtained from course management systems, classroom technology usage, or controlled online learning environments to inform educational decision making. However to the best of knowledge there is no research found to directly mine and analyse student-posted content from uncontrolled spaces on the social web with the clear goal of understanding students learning experiences.

II. LITERATURE SURVEY

A. *eMUSE*

It provides integrated access to all the Web 2.0 tools selected by the instructor for the course at hand: common access point, detailed usage instructions, summary of the latest activity. It retrieve students' actions with each tool and store them in a local database and offer a summary of each student's activity, including graphical visualization, evolution over time, comparisons with peers, as well as aggregated data eMUSE compute a score based on the recorded student activity (following instructor-defined criteria) and provide basic administrative services (authentication service, enroll students to the course, edit profile etc.). The fact that students have a place where they can access their own accounts to all Web 2.0 tools required for the course, as well as the accounts of their peers, provides an ease of access as well as a reduction in the time and effort needed for the tool management task. Another advantage of eMUSE is that it integrates Web 2.0 tools that learners are already familiar with from out-of-school activities, like Blogger, MediaWiki, Twitter, Delicious, YouTube etc. Thus, students have the opportunity to use the pedagogically valuable tools in a semi-formal framework, in collaboration with their peers, inside the eMUSE platform. In this sense, eMUSE is somewhat similar to Personal Learning Environments, like MUPPLE.

The first step towards the creation of eMUSE was to select the most suitable Web 2.0 tools to be integrated into the system, which meet two requirements: i) have a demonstrated pedagogical value (according to case studies reported in the literature); ii) offer technical support for mashup integration (well documented and maintained APIs, RSS feeds etc.). The integration of the Web 2.0 tools into the platform was done by means of mashups, ensuring a lightweight architecture, with loosely-coupled components. A mashup represents a combination of data and/or functionalities from two or more external sources to create a new Web application[10].

B. *Enhancing Learning with Visualization Techniques*

Information visualization is a powerful means of making sense of this data that has emerged from research in human-computer interaction, computer science, graphics, visual design, psychology, and quantitative data analysis. It is a growing field that is increasingly applied as a critical component in scientific research, digital libraries, data mining, financial data analysis, market studies, manufacturing production control, and drug discovery. The use of visualization techniques in learning is not new. They have been used in maps and drawings for thousands of years. This analyses how more novel visualization techniques can be used to enhance various activities during the learning process: finding and understanding educational resources, collaboration with learners and teachers, (self-) reflecting about learners' progress, and designing learning experience.

Visualizations also play an import role in other learning domains such as mathematics where they enable students to see the unseen in data. It is hard to find any mathematics textbook that does not use visualization techniques for explaining mathematical concepts such as the Pythagorean theorem. Presmeg provided a thorough review of research on visualization in learning and teaching mathematics since 1980. Computer Supported Collaborative Learning (CSCL) where learning is not only a matter of accepting fixed facts, but it is the dynamic, on-going, and evolving result of complex interactions primarily taking place within communities of people. Visualization of a social network can therefore be extremely useful to make people aware of their social context and to enable them to explore context[6].

C. *MOOC*

Advanced educational technologies are developing rapidly and online MOOC courses are becoming more prevalent, creating an enthusiasm for the seemingly limitless data-driven possibilities to affect advances in learning and enhance the learning experience. For these possibilities to unfold, the expertise and collaboration of many specialists will be necessary to improve data collection, to foster the development of better predictive models, and to assure models are interpretable and actionable. The big data collected from MOOCs needs to be bigger, not in its height (number of students) but in its width—more meta-data and information on learners' cognitive and self-regulatory states needs to be collected in addition to correctness and completion rates. This more detailed articulation will help open up the black box approach to machine learning models where prediction is the primary goal. Instead, a data-driven learner model approach uses fine grain data that is conceived and developed from cognitive principles to build explanatory models with practical implications to improve student learning.

Using data-driven models to develop and improve educational materials is fundamentally different from the instructor-centered model. In data-driven modeling, course development and improvement is based on data-driven analysis of student difficulties and of the target expertise the course is meant to produce; it is not based on instructor self-reflection as found in purely instructor-centred models. To be sure, instructors can and should contribute to interpreting data and making course redesign decisions, but should ideally do so with support of cognitive psychology expertise. Course improvement in data-driven modelling is also based on course-embedded *in vivo* experiments (multiple instructional designs randomly assigned to students in natural course use, also called "A/B testing") that evaluate the effect of alternative course designs on robust learning outcomes. In courses based on cognitive science and data-driven modelling, student interaction is less focused on reading or listening to an instructor's delivery of knowledge, but is primarily about students' learning by example, by doing and by explaining. In addition to avoiding the pitfall of developing intera ctive activities that do not provide enough useful data to reveal student thinking, MOOC developers and data miners must avoid potential pitfalls in the analysis and use of data. One such pitfall is the application of sophisticated statistical and machine learning techniques to educational data without understanding or contributing to relevant cognitive a nd pedagogical principles. This "black box model" approach focuses on improving prediction wi thout regards to understanding what is happening c ognitively (i.e., inside the the box). Using data-driven learner models to improve courses contrasts with the instructor-centered model in three key ways. First, course development and improvement is based not solely on instructor self-reflection, but on a data-driven analysis of student difficulties and of the target expertise the course is meant to produce. Second, course improvement is based on course-embedded *in vivo* experiments that evaluate the effect of alternative course designs on robust learning outcomes. Third, course interaction is not centrally about instructor's delivery knowledge, but about student learning by example, by doing and by explaining.

D. Learning Analytics and Educational Data Mining

Learning Analytics and educational data mining are data-driven approaches emerging in education. These approaches analyse data generated in educational settings to understand students and their learning environments in order to inform institutional decision-making. First data analysed using these approach typically are structured data including administrative data, students activity and performance data from CMS (Course Management System). In prediction, the goal is to develop a model which can infer a single aspect of the data (the predicted variable, similar to dependent variables in traditional statistical analysis) from some combination of other aspects of the data (predictor variables, similar to independent variables in traditional statistical analysis). Structure discovery algorithms attempt to find structure in the data without an a priori idea of what should be found, a very different goal than in prediction. In prediction, there is a specific variable that the EDM/LA researcher attempts to model; by contrast, there is not a specific variable of interest in structure discovery. In relationship mining, the goal is to discover relationships between variables in a data set with a large number of variables. Broadly, there are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining, and causal data mining[1].

EDM and LA methods have similarly been useful in understanding student learning in various collaborative settings. Collaborative learning behaviors have been analyzed in order to determine which behaviors are characteristic of more successful groups and more successful learners, in multiple contexts, including computer-mediated discussions online collaboration using software development tools , and interactive table top collaboration.

E. Crowd Based Design Activities

By definition, human-centered design relies on interaction with users. While interacting with userswithin industry can be challenging, fostering these interactions in a classroom setting can be even more difficult. This qualitative study explores the use of crowd-based design activities as a way to support student-user interactions online. There is a growing demand for human-centered design instruction as industry and government look for new ways to prepare students for careers in innovation. Instructors teach students the importance of authentic user interactions as users can provide a better under standing of real-world needs, help generate useful and creative solutions, and provide useful feedback.

Typically, designers interact with and study users through in-person research methods, such as contextual inquiry, interviews, and user enactments. However, orchestration challenges, such as locating users and setting up meetings, can limit the opportunities for such interactions. While these methods offer a rich understanding of users, performing these tasks could take weeks or months. The Internet offers a supplementary approach to reaching potential users. Designers in industry and academia have already begun exploring the value of soliciting design feedback and ideas online, such as testing first impressions of web-pages through an online usability tool and using crowd funding platforms[5].

F. Text-based mood classification

Mood is a strong form of sentiment expression, conveying a state of the mind such as being happy, sad or angry. Social media texts are rich in sentiment and this describes various fundamental issues related to mood sensing from these texts and novel applications of this information. Text-based mood classification and clustering, as a sub-problem of opinion and sentiment mining, have many potential applications, as identified in, such as automated recommendation for product websites, as a sub-component of web technology in business and government intelligence, or for the collection of empirical evidence for studies in psychological and behavioural sciences. Specifically, in the blogosphere, mood classification can be used to filter search results, to ascertain the mental health of communities, or to gain detailed insight into patterns of how bloggers behave and relate to one another[8].

However, text-based mood analysis poses additional challenges beyond standard text categorisation and clustering. The complex cognitive pro cesses of mood formulation make it dependent on the specific social context of the user, their idiosyncratic associations of mood and vocabulary, their syntax and style which reflects on language usage (for example, the order of linguistic components) and the specific genre of the text. In the case of weblogs challenges are reflected in the diverse styles of expression of the bloggers, the relatively short text length and the use of informal language, such as jargon, abbreviations and non-standard grammar. Feature-selection methods available in machine learning are often computationally expensive, relying on labelled data to learn discriminative features. However, the blogosphere is vast and is continuing to grow, making it desirable to construct a feature set that works without requiring supervised feature training to classify mood. To this end, it is necessary to lo ok to the results of studies that intersect Psychology and Linguistics.

G. Mining Twitter Data

Researchers from diverse fields have analyzed Twitter content to generate specific knowledge for their respective subject domains. For example, Gaffney analyzes tweets with hashtag #iranElection using histograms, user networks, and frequencies of top keywords to quantify online activism. Analysis methods used usually includes qualitative content analysis, linguistic analysis, network analysis, and some simplistic methods such as word clouds and histograms. The classification model based on inductive content analysis applied and validated on dataset. Thererfore, it emphasize not only the insights gained from dataset, but also the application of the classification algorithm to other datasets for detecting students problems. The human effort is thus augmented with large-scale data analysis. Popular classification algorithms include naive bayes, decision tree, logistic regression, maximum entropy, boosting and support vector machine. Based on the number of classes involved in the classification algorithms, there are binary classification and multi-class classification approaches. In binary classification, there are only two classes, while multiclass classification involve more than two classes. Both binary classification and multiclass classification are single-label classification systems. Single-label classification means each data point can only fall into one class where all classes are mutually exclusive[1].

Most existing studies found on tweet classification are either binary classification on relevant and irrelevant content, or multi-class classification on generic classes such as news, events, opinions, deals, and private messages. Sentiment analysis is another very popular three-class classification on positive, negative, or neutral emotions/opinions. Sentiment analysis is very useful for mining customer opinions on products or companies through their reviews or online posts. It finds wide adoption in marketing and customer relationship management. Many methods have been developed to mine sentiment from texts. For example, both Davidov et al. and Bhayani et al. use emotions as indicators to provide noisy labels to the tweets thus to minimize human effort. However, only knowing the sentiment of student-posted tweets does not provide much actionable knowledge on relevant interventions and services for students. The purpose is to achieve deeper and finer understanding of students experiences especially their learning-related issues and problems. To determine what student problems a tweet indicates is a more complicated task than to determine the sentiment of a tweet even for a human judge[21]. Therefore, it requires a qualitative analysis and is impossible to do in a fully unsupervised way.Multilabel classification, however, allows each data point to fall into several classes at the same time.

H. Text Pre-processing

Twitter users use some special symbols to convey certain meaning. For example, # is used to indicate a hashtag, RT is used to indicate a re-tweet. Twitter users sometimes repeat letters in words so that to emphasize the words , for example huuuuungryyy, sooo muuchh and monndayyy[1]. Besides, common stopwords such as a, an, and, of, she, it, non letter symbols, and punctuation also bring noise to the text. So preprocessed the texts before training the classfier:

1) First removed all the #engineeringProblems hashtags. For other co-occurring hashtags, only # sign is removed, and kept the hashtag texts.
2) Negative words are useful for detecting negative emotion and issues. So it substituted words ending with \n't and other common negative words(nothing, never, none, cannot) as\negtoken.
3) Then all words that contain non-letter symbols and punctuation are removed. This @ included the removal of a and http links, The RTs are also removed.
4) For repeating letters in words,our strategy was that when it detect two identical letters repeating then both of them are kept. If user detect more than two identical letters repeating, then it isreplaced with one letter. Therefore, huuuungryyy and sooo were corrected to hungry and so. muuchh was kept as muuchh. Original correct words such as too and sleep were kept as they were.
5) IT used the lemur information retrieval to remove the common stopwords. Then kept words like "much, more, all, always, still, only", because the tweets frequently use these words to express extent.

I. Use of web mining in studying innovation

As enterprises expand and post increasing information about their business activities on their websites, website data promises to be a valuable source for investigating innovation. This article examines the practicalities and effectiveness of web mining as a research method for innovation studies. We use web mining to explore the R&D activities of 296 UK-based green goods small and mid-size enterprise. The website data offers additional insights when compared with other traditional unobtrusive research methods, such as patent and publication analysis. It examine the strengths and limitations of enterprise innovation web mining in terms of a wide range of data quality dimensions, including accuracy, completeness, currency, quantity, exibility and accessibility[20]. While traditional methods offer information about the early phases of R&D and invention through publications and patents, web mining offers insights that are more downstream in the innovation process. Handling website data is not as easy as alternative data sources, and care needs to be taken in executing search strategies. Website information is also self-reported and companies may vary in their motivations for posting (or not posting) information about their activities on websites. Nonetheless, web mining is a significant and useful complement to current methods, as well as offering novel insights not easily obtained from other unobtrusive sources.

III. CONCLUSION

Students have the opportunity to use the pedagogically valuable tools in a semi-formal framework, in collaboration with their peers, inside the eMUSE platform. In this sense, eMUSE is somewhat similar to Personal Learning

Environments, like MUPPLE. In data-driven modeling, course development and improvement is based on data-driven analysis of student difficulties and of the target expertise the course is meant to produce; it is not based on instructor self-reflection as found in purely instructor-centered models. Visualizations also play an import role in other learning domains such as mathematics where they enable students

to see the unseen in data. EDM and LA methods have similarly been useful in understanding student learning in various collaborative settings. . Social media texts are rich in sentiment and this describes various fundamental issues related to mood sensing from these texts and novel applications of this information. Web mining is a significant and useful complement to current methods, as well as offering novel insights not easily obtained from other unobtrusive sources.

ACKNOWLEDGMENT

The authors would like to thank the publishers and researchers for making their resources available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

REFERENCES

[1] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan , “Mining Social Media Data for Understanding Students’ Learning Experiences”, IEEE Transactionon Learning Technologies, 2014.

[2] Carlos Castillo, Mohammed El-Haddad, Jurgen Pfeffer and Matt Stempeck, “Characterizing the life cycle of online news stories using social media reactions”, ACM conference on Computer supported cooperative work and social computing, 2014.

[3] Yue Gao, Fanglin Wang, Huanbo Luan and Tat-Sang Chau, “Brand data gathering from live social media streams”, ACM, 2014.

[4] Abdullah Gok, Alec Waterworth and Philip Shapira, “Use of web mining In studying innovation”, Scientometrics, 2014.

[5] Julie S Hui, Elizabeth M Gerber and Steven P Dow, “Crowd-based design activities helping students connect with users online”, ACM, 2014.

[6] Joris Klerkx, Katrien Verbert and Erik Duval, “Enhancing Learning with Visualization techniques”, In Handbook of Research on Educational Communications and Technology, Springer, 2014.

[7] Fei Liu, Maria Vasardani and Timothy Baldwin, “Automatic identification of locative expressions from social media text: A comparative analysis”, In Proceedings of the 4th International workshop on Location and the Web, ACM, 2014.

[8] Thin Nguyen, Dinh Phung, Brett Adams and Svetha Venkatesh, “Mood sensing from social media texts and its applications”, Knowledge and
Information systems, 2014.

[9] Marek Opuszko and Johannes Ruhland, “Classification analysis in Complex online social networks using semantic web technologies”, In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, IEEE Computer Society, 2012.

[10] Elvira Popescu, “Providing collaborative learning support with Social Media in an integrated environment”, World Wide Web,Springer, 2014

[11] Reihanch Rabbany, Samira Elatia, Mansoureh Takaffoli and Osmar R Zaiane, “Collaborative learning of students in online discussion forums:A social network analysis perspective”, In Educational Data Mining, Springer2014.

[12] William Ribarsky, Derek Xiaoyu Wang, and Wenwen Dou, “Social Media Analytics for competitive Advantage”, Elsvier, 2014.

[13] Cristobal Romero and Sebastian Ventura, “Educational Data Mining : A Review of the state of the art”, IEEE Transactions, 2010.

[14] Saajan Shridhar, Ankur Gupta and Swapan Shridhar, “Improving student Engagement in higher education: An experiment with a Facebook Application in India”, International Journal of Computer Science: Theory, Technology and Applications(IJCS), 2014.

[15] Jie Tang, Yaun Zhang, Jimeug Sun, Jinhai Rao, Wenjing Yu, Yiran Chen and Alvis Cheuk M Fong, “Quantitative study of individual emotional states in social networks”, IEEE Transactions, 2012.

[16] Christophe Thovex and Francky Trichet, “Opinion mining and semantic Analysis of touristic social networks”, In Advances in Social Networks Analysis and Mining , 2013 IEEE/ACM International Conference.

[17] Suppawong Tuarob, Conrad S Tucker, Marcel Salathe and NilamRam, “An ensemble heterogeneous classification methodology for discovering Health-related knowledge in social media messages”, Journal of Biomedical informatics, 2014.

[18] Suwimon Vongsingthong and Nawaporn Wisitpongphan, “Classification of university students’ behaviors in sharing information on Facebook”, Computer Science and Software Engineering(JCSSE), 2014 11th International Joint Conference, IEEE, 2014.

[19] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, “Data Mining with Big Data”, Knowledge and Data Engineering, IEEE Transactions, 2014.

[20] M.Vorvoreanu, Xin Chen and K.Madhavan, “ A Web-based tool for Collaborative social media data analysis”, International Conference on Social computing and its Application, 2013

[21] Andrei Yakushev and Sergey Mityagin, “Social networks mining for Analysis and Modeling Drugs Usage”, Elsevier, 2014.