Mining Social Media Data using Naïve Bayes Algorithm

Swapnaja Suryawanshi
Ekta Pawar
Pranali Shendekar
Kajal Lokhande
Apeksha Mengade

ABSTRACT
Online services provide a range of opportunities for understanding human behavior through the large aggregate data sets that there operation collects. Social network services have become a viable source of information for users. Studying the characteristics of such popular message is important for a number of tasks such as, breaking news detection, personalized message recommendation, others. We formulate the task into the classification problem and study two of its variants by investigating a wide spectrum of features based on the contents of messages, temporal information, metadata of messages and users, as well as structural properties of the user's social graph on large scale dataset. Students' informal conversations on social media shed light into their educational experiences opinions, feelings, and concerns about the learning process. Data from such 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. In this paper, we developed a workflow to integrate both qualitative analysis and large-scale data mining techniques. We focused on engineering students' Twitter posts to understand issues and problems in their educational experiences.

Keywords
Social media, Classification, Educations, Computer and educations

1. DEFINITION
In our project, we are implementing a multi label classification model where we are allowing one post to fall into multiple categories at the same time. Our classification is also at a finer granularity compared with other generic classifications. Our work extends the scope of data driven approaches in education such as learning analytics and educational data mining.

2. BACKGROUND
Social network services have become important communication tools for many online users. Such websites are increasingly used for communicating breaking news, eyewitness accounts and organizing group of people. Users of the services have become accustomed to receiving timely updates on important events, both of personal and global importance. Since our online behavior is in variably tracked, data sets of online activity can be used to make range of inferences about behavior, both in term of service itself and our broader lives.

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 obscured over time.

The research goals of this study 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.

We chose to focus on engineering students' posts on Twitter about problems in their educational experiences mainly because:

1. Engineering schools and departments have long been struggling with student recruitment and retention issues. Engineering graduates constitute a significant part of the nation's future workforce and have a direct impact on the nation's economic growth and global competency.

2. Based on understanding of issues and problems in students' life, policymakers and educators can make more informed decisions on proper interventions and services that can help student's overcome barriers in learning.

3. Twitter is a popular social media site. Its content is mostly public and very concise (no more than 140 characters per tweet). Twitter provides free APIs that can be used to stream data. Therefore, we chose to start from analyzing students' posts on Twitter.

3. RELATED WORK
The theoretical foundation for the value of informal data on the web can be drawn from Goffman's theory of social performance [1]. Although developed to explain face-toface interactions, Goffman's theory of social performance is widely used to explain mediated interactions on the web today [2]. Many studies show that social media users may purposefully manage their online identity to "look better" than in real life [3], [4].

Other studies show that there is a lack of awareness about managing online identity among college students [5], and that young people usually regard social media as their personal space to hang out with peers outside the sight of parents and teachers [6]. Researchers from diverse fields have analyzed Twitter content to generate specific knowledge for their respective subject domains. For example, Gaffney [7] analyzes tweets with hashtag #iranElection using histograms, user networks, and frequencies of top keywords to quantify online activism. Similar studies have been conducted in other

fields including healthcare [8], marketing [9], athletics [10], just to name a few.

4. ALGORITHM

Naïve Bayes Multi-label Classifier

One popular way to implement multi-label classifier is to transform the multi label classification problem into multiple single-label classification problems. One simple transformation method is called one-versus-all or binary relevance. The basic concept is to assume independence among categories, and train a binary classifier for each category. All kinds of binary classifier can be transformed to multi-label classifier using the one-versus-all heuristic. The following are the basic procedures of the multi-label Naïve Bayes classifier. Suppose there are a total number of N words in the training document collection (in our case, each tweet is a document) W = $\{w_1, w_{2,\dots,} w_N\}$ and a total number of L categories C=$\{c_1, c_{2,\dots,} c_L\}$. If a word w_n appears in a category c for $m_{w_n c}$ times, and appear in categories other than c for $m_{w_n c\prime}$ times, then based on the Maximum Likelihood Estimation, the probability of this word in a specific category c is

$$\mathrm{p}(w_n/c)=\frac{m_{w_n c}}{\sum_{n=1}^{N} m_{w_n c}} \quad (1)$$

Similarly, the probability of this word in categories other than c is

$$\mathrm{p}(w_n/c')=\frac{m_{w_n c\prime}}{\sum_{n=1}^{N} m_{w_n c\prime}} \quad (2)$$

Suppose there are a total number of M documents in the training set, and C of them are in category c. Then the probability of category c is

$$\mathrm{p}(c) = \frac{C}{M\prime} \quad (3)$$

And the probability of other categories c' is

$$\mathrm{p}(c') = \frac{M-C}{M} \quad (4)$$

For a document d_i in the testing set, there are K words $W_{d_i} = \{w_{i1,} w_{i2,} w \dots, w_{ik}\}$ and W_{d_i} is a subset of W. The purpose is to classify this document into category c or not c. We assume independence among each word in this document, and any word w_{ik} conditioned on c or c' follows multinomial distribution. Therefore, according to Bayes' Theorem, the probability that d_i belongs to category c is

$$p(c/d_i)=\frac{p\left(\frac{d_i}{c}\right).p(c\prime)}{p(d_i)} \prod_{k=}^{k} p\left(\frac{w_{ik}}{c}\right).p(c) \quad (5)$$

And the probability that d_i belongs to categories other than c is

$$p(c'/d_i)=\frac{p\left(\frac{d_i}{c\prime}\right).p(c\prime)}{p(d_i)} \alpha \prod_{k=}^{k} p\left(\frac{w_{ik}}{c\prime}\right).p(c') \quad (6)$$

Because $p(c/d_i) + p(c'/d_i)$=1, we normalize the latter two items which are proportional to $p(c/d_i)$ and $p(c'/d_i)$ to get the real values of $p(c/d_i)$ is larger than the probability threshold T, then d_i belongs to category c, otherwise, d_i does belong to category c. Then repeat this procedure for each category. In our implementation, if for a certain document, there is no category with a positive probability larger than T, we assign the one category with the largest probability to this document. In addition, "others" is an exclusive category. A tweet is only as- signed to "others" when "others" is the only category with probability larger than T.

5. CONCLUSION

Our study is beneficial in learning analytics, educational data mining, and learning technologies. It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content.

We provide great attention needs to be paid to protect students privacy when trying to provide good education and services to them.

Our study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering student's college experiences.

6. REFERENCES

[1] E. Goffman, The Presentation of Self in Everyday Life. Lightning Source Inc, 1959.

[2] E. Pearson, "All the World Wide Web's a Stage: The performance of identity in online social networks," First Monday, vol. 14, no. 3, pp. 1–7, 2009.

[3] J. M. DiMicco and D. R. Millen, "Identity management: multiple presentations of self in facebook," in Proceedings of the 2007 international ACM conference on Supporting group work, 2007, pp. 383–386.

[4] M. Vorvoreanu and Q. Clark, "Managing identity across social networks," in Poster session at the 2010 ACM Conference on Computer Supported Cooperative Work, 2010.

[5] M. Vorvoreanu, Q. M. Clark, and G. A. Boisvenue, "Online Identity Management Literacy for Engineering and Technology Students," Journal of Online Engineering Education, vol. 3, no. 1, 2012.

[6] M. Ito, H. Horst, M. Bittanti, danah boyd, B. Herr-Stephenson, P. G. Lange, S. Baumer, R. Cody, D. Mahendran, K. Martinez, D. Perkel, C. Sims, and L. Tripp, "Living and Learning with New Media: Summary of Findings from the Digital Youth Project," The John D. and Catherine T. MacAuthur Foundation, Nov. 2008.

[7] D. Gaffney, "#iranElection: Quantifying Online Activism," in WebSci10: Extending the Frontier of Society On-Line, Raleigh, NC, 2010.

[8] S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson, "'I can't get no sleep': Discussing #insomnia on Twitter," in Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems, 2012, pp. 1501–1510.

[9] M. J. Culnan, P. J. McHugh, and J. I. Zubillaga, "How large US companies can use Twitter and other social media to gain business value," MIS Quarterly Executive, vol. 9, no. 4, pp. 243–259, 2010.

[10] M. E. Hambrick, J. M. Simmons, G. P. Greenhalgh, and T. C. Greenwell, "Understanding professional athletes' use of Twitter: A content analysis of athlete tweets," International Journal of Sport Communication, vol. 3, no. 4, pp. 454–471, 2010.