Tracking Student Sentiment from Social Media

Dahai Guo
Department of Software Engineering, Florida Gulf Coast University, Fort Myers, FL, USA
Email: dguo@fgcu.edu

***Abstract*—This paper describes a framework that utilizes technologies to track students' sentiment using their input to social media. Social media has accumulated a vast amount of textual inputs from students. That amount is still growing rapidly. While it is desirable to gain insights from these inputs, it is impossible to manually analyze individual inputs. Fortunately, computer technologies exist for summarizing textual data. One of such technologies is referred to as sentiment analysis, which has been used in the business world for tracking customers' opinions on certain products or services. The framework, introduced in this paper uses these sentiment analysis technologies to track students' sentiment. It consists of three components: 1) data collector, 2) sentiment analyzer, and 3) result reporter. In addition, this paper also presents a case study where students' comments on ratemyprofessors.com are analyzed.**

***Index Terms*—sentiment analysis, information retrieval, social media**

I. Introduction

A. Sentiment Analysis

Sentiment Analysis refers to a set of computer algorithms which take textual documents as input and classify it as expressing a certain level of positive or negative opinions or sentiment [1]. A document can be any textual data, such as online posts, blogs, user reviews, etc. The sentiment analysis technologies may analyze it, taking into consideration issues like the frequencies of positive/negative words, the structure of the sentences, the subject which the document is related, etc. [1]. For example, a student review "This professor is very clear and helpful." could be classified as a document with a positive sentiment. Another review "The class is horrible, I hate that." could represent some negative sentiment. Apparently, the more the textual data, the more accurately the sentiment can be identified. There have been studies that address the accuracy of sentiment analysis. The algorithms found at Stanford University have achieved 85.4% accuracy. [2] Researchers at the University of Pittsburgh found an accuracy of 82%. [3]

In the business world, the sentiment analysis technologies have been used to measure and quantify customer opinions to help respond and engage to customer needs [4]. The article [5] by Kho has pointed out "*There's special business value in discerning opinion, sentiment and subjectivity—the 'voice of the customer'—in text as varied as blogs, forum postings, articles, e-mail and survey responses. That field of "customer experience analysis" applies sentiment analysis and other techniques to understand and help predict consumer behavior via text analysis coupled with analysis of customer transactions, profiles and demographics.*" Given the value of sentiment analysis, the software industry has built sentiment analysis products, such as IBM Social Sentiment Index [6], Google Prediction API [7], Alchemy API [8], Text-Processing API [9] etc.

B. Social Media

Social media has accumulated a vast amount of data. Much of them are textual data. In this research, we concern the collection of these data using computer programs, instead of manually collecting the data. To the best of the author's knowledge, there are three ways for programmatically collecting the data:

1) Using web application programming interfaces (API's)

Many social media sites, such as Facebook and Twitter open their data at their special web-based entry points, referred to as their web API's. With these web API's, computer programs can connect to these social media sites and download their data. While being free, the amount of data which can be downloaded in this way is limited.

2) Purchasing through a data vendor

There are companies which collect data from social media sites and sell them to downstream users. Gnip, Inc. is such a company. While data purchased from these company are not free, a much larger amount of data can be obtained.

3) Using a web crawler to collect data

Some social media sites do not open their data either through web API's or a data "retailer". In this situation, a computer program can be written to be a "robot" Internet surfer. This kind of "robot" surfers is often referred to as web crawlers which navigate the web sites following hyper-links. At each page, it downloads the interested data. The amount of data that can be downloaded depends on different applications. In theory, everything in the Internet can be downloaded. But this may not be needed.

In the case study in Section 3, we use the third method because the data source (ratemyprofessors.com) do not provide web API's or sell their data to any data "retailer".

II. Framework for Tracking Students' Sentiment

In this section, a framework is proposed to track students' sentiment using social media data. This framework consists of three components as follows:

Manuscript received February 15, 2015; accepted May 2, 2015.

doi: 10.12720/jait.6.2.80-83

1) Social media data collector

A computer program which downloads textual data from social media via one or multiple of the three ways described in Section 1.B.

2) Sentiment analyzer

Another computer program which takes advantage of some existing software tools to process the data collected in the previous step to find the sentiment, identified from the text.

3) Result reporter

This component reports the result to the end user in the most meaningful and friendly way.

Fig. 1 describes this framework with focus on the data flow.

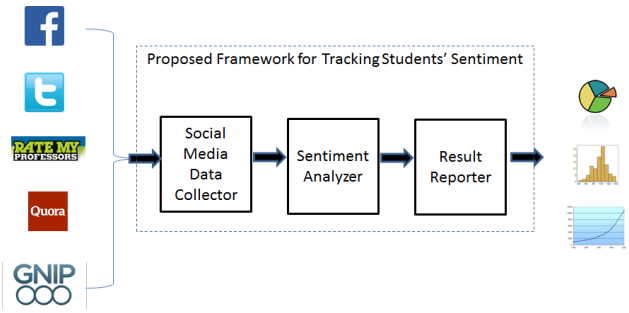

Figure 1. The proposed framework for tracking students' sentiment using social media data

A. Collecting Social Media Data

As described previously, there are three ways for downloading social media data in a computer program. The selection among them should depend on the available resources and the nature of the social media site. Ideally, the wanted data can be purchased from a data vendor, through which a large amount of data can be obtained. Of course, this may require a significant monetary investment to pay for the data. In addition, how to store these data can be another challenge considering the vast amount of data in social media. For a small scale research, using the web API's may be a good choice. However, if the wanted data are only accessible from web pages, the only option is to develop a web crawler to automatically visit the web pages and extracted the wanted text. A web crawler needs the developers to analyze the web pages where the interested data are and write correct algorithms. Given the fact that web pages are often dynamic, web crawlers may need to be updated periodically.

B. Performing Sentiment Analysis

Sentiment analysis involves a set of sophisticated computer algorithms. Fortunately, software tools have been developed to implement these algorithms. Here are several such software tools: 1) Stanford Natural Language Processing library (open-source) [2], 2) Text-Processing API [9] which requires some payment based on usage ($75 for analyzing 2.5 million characters), 3) IBM Social Sentiment Index [6] which is commercial software.

When performing the analysis, some preprocessing may be needed. For example, if a document includes textual data regarding to multiple subjects, it makes sense to split it to multiple more coherent documents. In addition, when using a web crawler to download social media data, the preprocessing could remove unrelated data, such as dates, authors, advertisement on a web page, etc.

C. Reporting the Results

This part of the framework is open-ended. It requires some domain knowledge. The author would make two suggestions. First, the outcome of the sentiment analysis is to assist in the decision making process. The technicality of the report should be much weakened because the audience does not need to understand how data is collected and/or analyzed. Secondly, the report needs to let the audience be aware that the accuracy of sentiment analysis algorithms, while being relatively high is not 100%. The report should avoid making assertive conclusions. Instead it should encourage follow-up investigation. In Section 3.C, an example will be demonstrated.

III. Case Study: Analyzing Data RateMyProfessors.com

This case study is based on the data from ratemyprofessors.com (RMP), which is a review site, primarily used by students in North America. It allows college students to assign rating to their professors. When a student is rating a professor, he/she must rate the professor in the categories: "easiness", "helpfulness", and "clarity" on a 1-5 scale. RMP also calculates the average of each professor's "helpfulness" and "clarity" to a score, referred to as "Overall Quality". Many students also would write some comments to explain the rating. It is noteworthy that RMP was recognized as one of the 50 best web sites of 2008.

A. Web Crawling

Unfortunately, ratemyprofessors.com (RMP) neither opens its data via web Application Programming Interfaces (web API's) nor makes their data available for purchasing through a data "retailer" like Gnip Inc. [10] The only option is to write a dedicated web crawler. This require us first to understand how RMP organizes their web pages and how data are presented in each page. This process can be time consuming and error-prone. Our web crawler, based on crawler4j [2] started to download student comments on each instructor and their average "Overall Quality" score, also inputted by students. Note it may take hours or even days to complete downloading all the data.

In this study, the web crawler collected student rating from RMP for 24,444 instructors in the United States. All the student comments add up to 3.7 million words.

B. Sentiment Analysis

The software used in this study is the Text-Processing API [9]. This tool does not need installation. Instead a computer program is needed to access this tool using its web application programming interfaces (web API's). We

chose this tool because of its relatively low cost ($75 for analyzing 2.5 million characters) and popularity. Text-Processing API is based on the Natural Language Toolkit which has been used a lot in education and research. [12]

The downloaded data are first split to files, each of which contains the collection of student comments on a specific instructor. Then each file is submitted to the software tool which then returns a score from [1-5], where 1 indicates the most negative sentiment and 5 indicates the most positive sentiment.

C. Data Analysis

After the sentiment analysis step, each instructor had a score, indicating his/her students' sentiment. As explained in 3.2, this sentiment score is in [1, 5]. In addition, each instructor at ratemyprofessors.com (RMP) has a numerical score, called "Overall Quality" also in the range [1], [5]. In this case study, we perform two types of analyses as follows:

1. For all the instructors in each major, average their student sentiment scores and "Overall Quality" scores.
2. For all the instructors in each state in the U.S, average their student sentiment scores and "Overall Quality" scores.

Fig. 2 and Fig. 3 demonstrate the average ratings with respect to different majors/states.

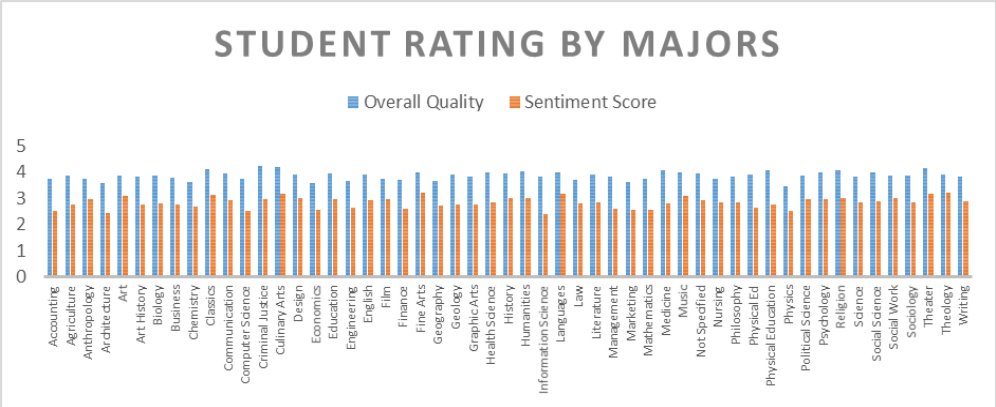

Figure 2. Average "overall quality" and "sentiment score" for different majors

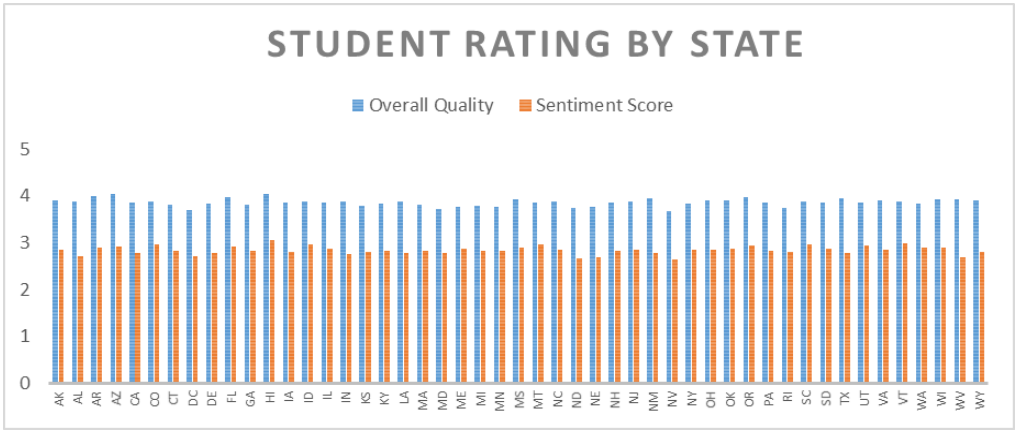

Figure 3. Average "overall quality" and "sentiment score" for different states in the U.S.

In Fig. 2 and Fig. 3, the average sentiment scores are significantly lower than the average "Overall Quality" scores. However, the higher the average sentiment score, the higher the overall quality score. In fact, the average difference is 1.02 with a standard deviation 0.08 on the 5-point scale. The difference between the sentiment scores and the "Overall Quality" scores appears to be very consistent. The difference may be explained by that students may tend to be more lenient when explicitly rating their instructors than when leaving comments.

Table I demonstrates the ten States where the highest and lowest student sentiment scores were found. Table II demonstrates the ten majors where the highest and lowest student sentiment scores were found.

TABLE I. TEN STATES WITH THE HIGHEST/LOWEST STUDENT SENTIMENT SCORES

Ten States with the Most Positive Sentiment	Ten States with the Most Negative Sentiment
Hawaii	Nevada
Virginia	North Dakota
South Carolina	West Virginia
Colorado	Nebraska
Idaho	Alabama
Montana	Washington D.C.
Oregon	Indiana
Utah	New Mexico
Arizona	Louisiana
Florida	Maryland

TABLE II. TEN MAJORS WITH THE HIGHEST/LOWEST STUDENT SENTIMENT SCORES

Ten Majors with the Most Positive Sentiment	Ten Majors with the Most Negative Sentiment
Fine Arts	Information Science
Theology	Architecture
Culinary Arts	Physics
Languages	Accounting
Theatre	Computer Science
Classics	Mathematics
Music	Marketing
Art	Economics
Humanities	Management
Religion	Finance

From Table II, students appear to have more positive sentiment in liberal art majors, while more negative sentiment is found in science and business related majors.

IV. CONCLUSIONS

In this study, a framework for tracking students' sentiment is proposed. This framework consists of three components: 1) data collector, 2) sentiment analyzer, and 3) result reporter. To collect data, there exist three ways: 1) using web API's, 2) purchasing data from data "retailers", and 3) develop a web crawler to automatically download the wanted data. This article also discusses the issues related to the sentiment analyzer and result reporter, such as preprocessing and effective decision making based on sentiment analysis.

A case study is presented in this article. It utilizes student comments from ratemyprofessors.com (RMP) to analyze their sentiment. It is found that the student sentiment demonstrates similar trend to the "Overall Quality" for their instructors. Also the analysis finds that more positive sentiment in more liberal arts related major and more negative one in majors where mathematics plays a more important role.

The case study suggests that the framework can be effective in tracking students' sentiment. Implementations of this framework can of course choose other data social media, such as Twitter, Facebook, etc. They can also use more software tools in analyzing sentiment so that the results can cross-examined. Lastly, the subject on which the sentiment is to be analyzed can be other topics, such as student activities, job placement, etc.

REFERENCES

[1] B. Liu, *Web Data Mining–Exploring Hyperlinks, Contents, and Usage Data*, 2nd ed. ch. 11 2011.

[2] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment Treebank," in *Proc. Conference on Empirical Methods in Natural Language Processing*, Seattle, USA, October 18–21, 2013.

[3] J. Wiebe and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 1. no. 2, December 2005.

[4] S. Grimes. (April 22, 2014). Sentiment Analysis and Business Sense. Clarabridge Blog. [Online]. Available: http://clarabridge.com/2014/04/sentiment-analysis-and-business-sense/

[5] N. Kho, "Customer experience and sentiment analysis," *Knowledge Management*, vol. 19, no. 2, February, 2010.

[6] IBM Social Sentiment Index. (2014). [Online]. Available: http://www.ibm.com/analytics/in/en/conversations/social-sentiment.html

[7] Google Prediction API. (2014). [Online]. Available: https://cloud.google.com/prediction/docs

[8] Alchemy API. (2014). [Online]. Available: http://www.alchemyapi.com/

[9] Text-Processing API. (2014). [Online]. Available: http://text-processing.com/

[10] Gnip Inc. (2014). [Online]. Available: http://www.gnip.com

[11] craw4j. (2014). [Online]. Available: https://code.google.com/p/crawler4j/

[12] B. Steven, E. Klein, E. Loper, and J. Baldridge, "Multidisciplinary instruction with the Natural Language Toolkit," in *Proc. the Third Workshop on Issues in Teaching Computational Linguistics*, ACL, 2008.

Dahai Guo has been on the faculty of Software Engineering at Florida Gulf Coast University since 2006. Currently, he is the Department Chair of Software Engineering. He previously was a visiting lecturer in the Department of Computer Engineering at the University of Central Florida (UCF) from 2005-2006. Dr. Guo received his B.S. in Computer Science and Applications from Shanghai Jiaotong University in China in 1999 and his M.S. and Ph.D. in Computer Engineering from UCF in 2001 and 2005, respectively. His currently research interests include social media data analysis, distributed systems, and virtual simulation. He has published several journal articles and presented at many national and international conferences.