Proceedings of Engineering and Technology Innovation, vol. 7, 2017, pp. 31 - 36

Using Text Mining to Extract Issues for School: an Empirical Study of the Social Platform-Dcard

Hsin-Yi Wang, Shu-Fen Chiou*, Jung-Wen Lo

Department of Information Management, National Taichung University of Science & Technology, Taichung, Taiwan, ROC.

Received 18 July 2017; received in revised form 15 August 2017; accepted 30 August 2017

Abstract

Nowadays, social network within sentiment analysis has become the main trend in text mining domain. There are many platforms have been analyzed, such as Facebook, Twitter, Instagram, and so on. In our manuscript, we attempt to extract the information about the sentiment polarity of messages (positive, neutral or negative) in a social platform "Dcard". The users of Dcard are Taiwanese college students, and anonymous post is being used this in social platform, therefore, the user can express their opinion more freedom. We use Dcard to the sentiment polarity of messages in extract the information about the school; moreover, the school could get the feedback from this finding to improve their policy. In this paper, we used python to scrap the web page, and the sentiment lexicon would be built.

Keywords: text mining, big data, social platform, sentiment

1. Introduction

Nowadays, Internet is used for communication widely. People used the internet to be browsing the web and collecting data (79.9%), using community websites (24.1%), playing online games (19.4%), listening to music or watching movies (17.8%), (15%) shopping online (10.3%) [1]. For community websites, people prefer communicating via the Internet Services over talking face-to-face or writing letters. They are more often writing blogs or posting messages on social networks and the personality will be presented by habitual vocabularies they used. In this research, we try to s analyzed Chinese vocabularies on a social platform named Dcard [2].

With the accumulation of large amounts of educational data, the use of advanced statistical techniques, such as data mining, exploring the potentially useful information or realizing some knowledge from bunch of data [3]. Another important material for structured data is unstructured data composed of free text; for example, scientific research papers, patented technical documents, qualitative interview data, and open questionnaires the content of the analysis, etc. with the analysis or research value. Analysis of these unstructured data exploration methods, relying on the free text of the advanced processing and statistical operation s, that is, text mining technology[4].

Dcard anonymous posting mechanism has successfully attracted tens of thousands of domestic and foreign college students to become platform members, this research is using Dcard community behavior of the Chinese corpus, the part of the analysis from the Dcard text behavior can be analyzed everyone Comment on the community on the web.

* Corresponding author. E-mail address: fen057@nutc.edu.tw

Tel.: +886-4-22196391

In this paper, we use the word surveying technology to analyze the administrative quality of the campus in order to avoid the loss of the students, but also to attract the resources and improve the quality of education. The results data can be used as reference for the follow-up research reference and campus administrators.

2. Related Works

In this section, we first introduce the technique which we use to extract Chinese vocabularies. And we describe the social platform named Dcard

2.1. Python

Python is a widely used high-level programming language for general-purpose programming, created by Guido van Rossum and first released in 1991. An interpreted language, Python has a design philosophy which emphasizes code readability (notably using whitespace indentation to delimit code blocks rather than curly braces or keywords), and a syntax which allows programmers to express concepts in fewer lines of code than possible in languages such as C++ or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale [5].

2.2. Dcard

Dcard is the first exclusive college student dating site platform in Taiwan. The official website is show in Fig. 1 [6]. The Dcard platform attracts many college students to join to expand their circle of friends, but to join must provide school e-mail, with the real name and photos through the administrator can be certified. Dcard using anonymous system to register, many students published articles to take anonymous system to express their own emotions, and these acts through the text presented in the article.

Fig. 1 Dcard website

2.3. Text mining

In a large number of data, there are digital structure of the structured data and text, sound, image of the unstructured data. There are already hundreds of ways to deal with structured data. [7]

Text mining from biological literature is emerging as one of the main issues in bioinformatics research, and NLP methods are regarded as being useful to raise the potential of text mining from this literature. While the techniques are separated relatively into domain-portable, reference materials, except for Corpora. [8].

Text mining is a new field, the text of the survey can be collected from the text of meaningful messages, for the file can also be analysed out of his specific purpose of the target, the text is undiscovered, invisible, and difficult in the algorithm. However, in modern culture, the document is the most common tool for the formal exchange of messages. The field of textual exploration usually involves a document whose function is to convey a factual message or opinion. [9]

The steps of knowledge exploration: [10]

- Data collection
- Data cleansing
- Data conversion
- Application of exploration technology
- The results are presented and interpreted

The methods of knowledge exploration:

- Association analysis
- Classification
- Clustering
- Summarization
- Prediction
- Sequence analysis

3. Our Proposed Method

The work presented in this article covers, on the one hand, the extraction of information about user's positive/neutral/negative sentiments from the text title they write. We investigate the detection of sentiment changes with respect to the "usual" sentiment of each user [11]. In Fig. 2, we present the methods' flowchart with each purpose below:

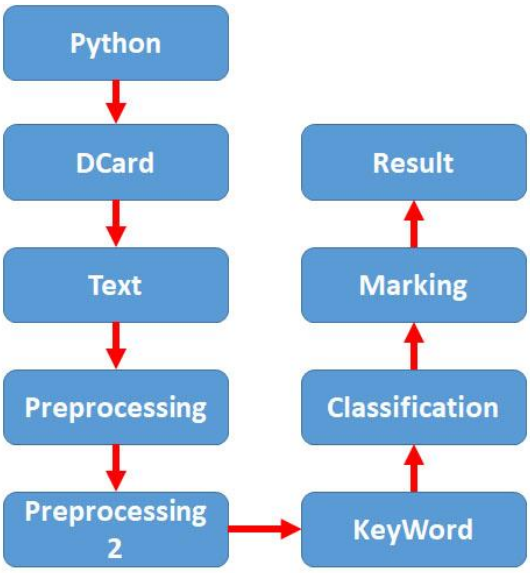

Fig. 2 Method

Table 1 Marking

Positive	其實我們的學校真的挺不錯的(Actually our school is really good)
Negative	畢業證書之學校爛行政(Graduation certificate of the school rotten administration)
Neutral	問學校語言學習軟體(Ask for school language learning software)

Table 2 Dcard to word

Web	Text
	我是多媒人我好美(不 中科大 多媒體設計學系 5 2 財金人 哈哈中科大終於開放囉!! 有沒有財金的朋友阿~~ 來這留下你們的腳印吧^^ 中科大 財務金融學系 7 4 中科大會資人 會資人來留下你們的足跡吧~~~ 剛剛進來Dcard時都去看別人發的文章 忘記 中科大 會計資訊學系 34 18

Table 3 Preprocessing 2

Before the process	After the process
【沒朋友 徵室友，我又來了】 (【No friends levy roommate, I have come】)	沒朋友徵室友我又來了 (No friends levy roommate, I have come)

Table 4 Keywords

Main Classification	Keyword		
Course	課程	課名	老師
Administration	足球隊(Football team)	行政(Administrative)	校園(Campus)

Step 1: We need data collection and marking. This research using the article title to classification to Positive, Neutral and Negative includes school course and administration tow part. We give an example in Table 1.

Step 2: We use Python to extract Dcard website to the text content. We extract Dcard website by using Python API URL. We extract Dcard website convert to text content, combine and export to document. Table 2 are our examples.

Step 3: Sorting out the data we export, and remove the extra number and English letter (the web content which is not created by the Dcard user), and convert the full shape character to half shape character.

Step 4: Remove the punctuation in the data (e.g. #, semicolon, comma and space) like Table 3.

Step 5: Create Keywords. Classification the pre-process data we created for the school course and administration. For an example in Table 4, we set the course and its name for the school course partition, and the “soccer team” for administration due to the students usually use

4. Data Analysis and Results

4.1. Data

We collect the Dcard information from 2014 to 2017 in Table 5. The results of the original data have number of 24939. The data have been 4941 after pre-process finishing. We the separate these into "Administration", "Courses" and "Other" three classifications:

Table 5 Data

Main Classification	Data
Administration	1098
Courses	729
Other	3114

Table 6 Keywords dictionary

Main Classification	Keywords					
Administration	行政 (Administrative)	足球隊 (Football team)	停車場 (Parking lot)	學校 (School)	校園 (Campus)	系學會 (Department of Science)
	智慧大師 (Master of wisdom)					
Courses	課 (Class)	課程名稱 (Class name)	老師名稱 (Teacher)	RS (RS)	多益 (TOEIC)	科系 (Department)

Table 7 Polarity

	Administration	Courses
Positive	251	211
Negative	321	215
Neutral	526	303

4.2. Keyword dictionary

In the classification keywords shown in Table 6, we found that students had certain classification of keywords. For instance, students often use the "足球隊" to describe the attitude of the campus administration. The following keywords which are usually being used:

4.3. Data polarity

We will separate relatively into positive, negative and neutral, the following is a review of a variety of data classification number:

5. Discussion

In Fig. 3, studies have investigated the title Dcard reviews, we found that 22% of students expressed the article ideas for the administration, 15% of people have questions or ideas for courses in Dcard, and the others are expressed Other demands.

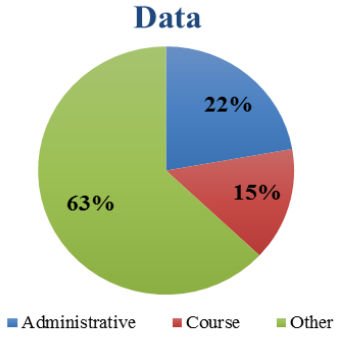

Fig. 3 Data

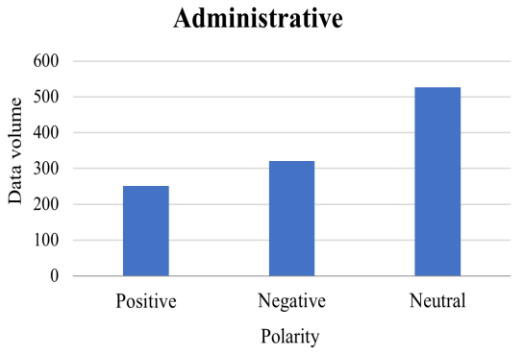

Fig. 4 Administrative

In Fig. 4, we also found that students in the comments can really achieve the administration and course, there are 1098 administrative from administration. 251 is the positive title, 321 negative title and 526 neutral title. 52 percent of the students published their perspective, point of view and opinions on the Dcard, 29% of whom expressed negative title, which showed that students were less likely to be administratively in school (Including school systems).

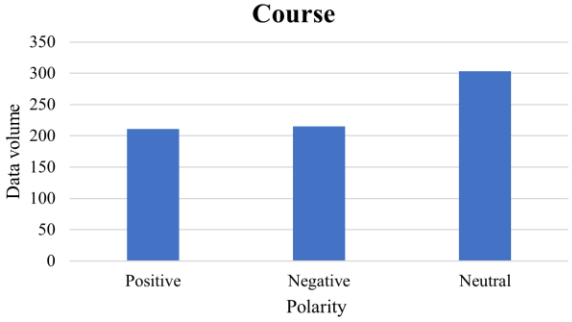

Fig. 5 Course

There are 729 titles in the course, with 211 positive titles, 215 negative titles and 303 neutral titles, which shows that students have 58% of the ideas and suggestions in the course or teacher, and the 42% of the people have other problem in course. We can see the results in Fig. 5.

6. Conclusions

This research explores students' perceptions of campus quality, extracts web pages, use keywords and the terminology of classified sentences to achieve the desired results. Use the positive and negative sentences to explore the ideas, perspectives, point of views of students on the administration and the curriculum the negative title, it can be extracted separately review and achieve to improvement the purpose campus administration and curriculum.

Furthermore, future research can be published on Dcard and use the text to explore the students for the campus administration ideas and recommendations make research more in-depth can improve the improvement. There are two limitations to this paper, which can be discussed in subsequent studies:

1. Research methods can only see the school administration is missing, but cannot see what kind of administrative issues are
2. Title cannot tell what kind of problem, is it can only know the advantages and disadvantages

References

[1] S. W. Wu, "Applying text mining technique to analyze the software quality characteristics of mobile games," In department of industrial management, National Pingtung University of Science and Technology, pp. 1-49, 2013.
[2] X. Z. Chang and Y. M. Li, "Using text mining to predict personality based on social behavior," 2012.
[3] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: a review and future visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3-17, 2009.
[4] Y. H. Tseng and Y. I. Lin, "The application of content mining techniques to the analysis of educational evaluation research trends," Journal of Research in Education Sciences, vol. 56, no. 1, pp. 129-166, 2011.
[5] G. van Rossum, Python, https://zh.wikipedia.org/wiki/Python#.E5.8F.82.E8.80.83.E6.96.87.E7.8C.AE, 1991.
[6] C. Y. Jian, "Dcard," https://www.dcard.tw/, 2011.
[7] C. M. Young, "Applies by text mining to the support systems for coding ICD-9-CM —a study of admission note and discharge summary," Taipei Medical University Information Management, pp. 1-68, 2004.
[8] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus—a semantically annotated corpus for bio-textmining," Bioinformatics, vol. 19, pp. i180-i182, 2003.
[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10-18, 2009.
[10] Y. H. Tseng, "Research and development on automatic information organization and subject analysis in recent decades," Journal of Educational Media and Library Sciences, vol. 51, pp. 3-26, 2014.
[11] A. Ortigosa, J. M. Martin, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," Computers in Human Behavior, vol. 31, pp. 527-541, 2014.