Sentiment Analysis: Determining People's Emotions in Facebook

N. AZMINA M. ZAMANI[1], SITI Z. Z. ABIDIN[1], NASIROH OMAR[1], M. Z. Z.ABIDEN[2]
[1]Faculty of Computer and Mathematical Sciences,
[2]Faculty of Architecture, Planning and Surveying,
Universiti Teknologi MARA
40450 Shah Alam, Selengor
MALAYSIA
namz.ina@gmail.com, {zaleha, nasiroh}@tmsk.uitm.edu.my, mzabiden009@gmail.com

Abstract: Social Network Sites (SNSs) play important roles in people's lives for sharing information. Facebook becomes one of the important platforms for interaction. Facebook allows people to have their own accounts to comment, express feelings and convey emotions via texts as well as emoticons. When a certain issue is discussed, monitoring such information becomes difficult since there are too many suggestions and the problems usually tend to be overlooked. Thus, this paper aims to identify the opinion mining and sentiment analysis components for extracting both English and Malay words in Facebook. Information, in terms of texts, are extracted and clustered into emotions. This work begins with transforming unstructured information into meaningful lexicons after extracting the Facebook's contents. All of the meaningful lexicons are stored in a database after manual identifications are carried out. With sentiment analysis, emotions are classified into happy (positive), unhappy (negative) and emotionless. The results are displayed by giving the percentage of sentiment categories so that it can be concluded that a selected Facebook post get positive or negative responses based on all comments received from users. As a case study, an issue on an examination results is posted and results of students' responses are determined. This study is significant of enabling the stakeholders such as administrators and businessmen to monitor any discussion issue for enhancing their services.

Key-Words: Opinion Mining, Sentiment Analysis, Lexicon-based, Facebook, Social Network, Quantifying sentiments.

1 Introduction

Social Network Sites (SNSs) such as Facebook, Twitter, blog and online forum play important roles in our daily lives where people are allowed to communicate and share information with others [1][2]. Facebook is one of the most popular online SNS [3][4]. It contains comments posted by people where they can express their emotions and opinions via texts and emoticons.

While many people concern on the number of *likes* and continue to build up network of friendships, this paper proposes to quantify sentiments in Facebook by analyzing the users' emotions toward an issue based on the comments in a post. The emotion (happy, unhappy and emotionless) patterns will quantify the sentiments. This work focuses on a randomly selected post from the university's (Universiti Teknologi MARA, Malaysia) *Student Financial Section* (Bahagian Pengurusan Kewangan Pelajar UiTM) Facebook page by analyzing both English and Malay texts in the comments, which are extracted and evaluated interactively. A software tool is developed by JAVA language to perform the sentiment classification by using Lexicon-based approach. A predefined list of words is stored in a database as corpus for later classification. Then, results are displayed based on percentages per category. The tool can allow stakeholders such as administrators and businessmen to monitor the issues so that the related people are able to enhance their services.

This paper is organized as follows. Section 2 discusses on the related work, while Section 3 shows the research methodology. Results and discussions are found in Section 4 before the concluding remarks in Section 5.

2 Related Work

Most of the Internet users are using the SNSs for online interactions. Normally, the users share opinions, facts or issues based on their topic of interest without being at the same place and same time [5][6]. There are a lot of tools for opinion mining and sentiment analysis such as analysis of customers' product reviews, personality of SNSs' users and educational purpose. Sentiments are analyzed after all the opinions in comments or postings are extracted.

By analyzing people's sentiment, the emotions of the public toward a particular issue can be observed, experimented and quantified [7]. The complexities in conveyed texts cause an insufficiency to abide in the existing sentiment analysis studies which identify user behaviours as well as their state of minds [8]. Based on previous research, various methods are implemented in sentiment analysis which is done either manually, semi-automatically or fully automatically. In addition, sentiment analysis consists of several processes which include extraction and classification.

2.1 Extraction and Classification

There are various methods used in previous research for data extraction where methods in sentiment analysis have transformed moderately from manual paper-based survey to automatic computer-based system. Manual approach such as surveys provides limited questions and answers for the participants as well as time-consuming in conducting the surveys for data collection. Unlike automated systems, there are no/less limitations in collecting and analyzing data [7].

Some of the automated tools perform filtration immediately after the contents are extracted. For example, a research done on Twitter extracts relevant and eliminates irrelevant contents from tweets by using information retrieval techniques then filters the data after the extraction. The filtration includes emoticons replacements, upper and lower casing, and removal of stop words, repeated words as well as punctuations [2][9].

In order to classify the texts into emotions, various classifiers or methods are used after performing the filtration of data. In [6], sentiment analysis is done by using Sentiment Identification Algorithm which are Compositional Semantic Rule, Numeric Sentiment Identification, and Bag-of-Word and Rule-based. All these algorithms are used in Machine Learning Model which involves several classifiers such as Decision Tree (J48), Random Forest, Logistic Regression and Neural Network. There are also other classifiers being used by other researchers in their works such as Naïve Bayes, Support Vector Machine (SVM), Sequential Mining Optimization (SMO) and Maximum Entropy. Even though the approach is used to increase the accuracy of sentiment classification, the manual Lexicon-based approach is more efficient for data analysis than the automated classification methods because it is flexible, especially for multilingual texts. Mostly, automated classification methods are used for English texts and as for the other texts such as Malay, Thai, Chinese, German and Spanish used either manual or semi-automated approach for classifications.

Besides mining and analyzing texts in sentiment analysis, emoticons are also taken into account to classify the emotions that the people conveyed since the emoticons are able to support voice inflections, facial expressions as well as bodily gestures on the SNS [10]. Emotions can be classified into three sets of texts which are texts containing positive and negative emotions as well as texts which only state a fact or do not express any emotions [11].

Manual approach of Lexicon-based is the best method for classifying words and emoticons occurrences since the flexibility is suitable for multilingual texts. Other automatic methods are more suitable to be used on English texts as most research done on English texts used automated tools. In order to categorize and recognize emotions or sentiments in text, lexicon is important. In [12], a Chinese emotion lexicon based on the emotion corpus, Ren-CECps 1.0 is created to conduct an automatic method of classifying emotions or sentiments. Moreover, there is also some research that used a combination of the Machine Learning methods and the manual Lexicon-based approach. The combinations of the two methods are performed due to the advantage and disadvantage of each method. Machine Learning techniques have the benefit of having a high accuracy given a high-quality training corpus, but the drawback of this technique is the involvement of an extensive manual work in annotating data. On the other hand, the Lexicon-based approach allows many domains to have an ease of use but unfortunately, the approach unable to yield comparable accuracy to the Machine Language techniques [13].

2.2 Facebook

Based on several studies and observations, Facebook is the most flexible SNS compared to the other SNSs since the issues can be viewed widely. In addition, Facebook status messages are more concise than reviews and also less complicated to classify. This will lead to a better writing and portrayal of emotions is more precise [9]. Facebook also consists of one-to-many communication style by using the Facebook "Walls" [14]. There are also several researches done on Facebook by using various methods for different purposes.

Most research had carried out on the study of users' interactions, social capital, personality and behaviours on Facebook based on the usage and activities as shown in Table 2. Furthermore, Facebook is also being studied based on online communicative language as done in [15] and also on education purposes by using Content Analysis method in which the contents of the Facebook comments is manually assessed.

Table 2: Overview of research done on Facebook and the methods used

RESEARCH STUDIES	METHODS USED
Users' personality and behaviours.	Measuring Facebook usage.
Users' interactions in Facebook.	Machine Learning techniques.
Facebook usage by the users.	Surveys
Social capital in Facebook.	
Education-related use in Facebook	Content Analysis.

Based on previous research done on Facebook, sentiment analysis is performed mostly in English texts only and very few in Malay. In order to provide another alternative for sentiment analysis, this work focuses on quantifying Facebook sentiments by using Lexicon-based approach for both English and Malay texts.

3 Research Methodology

The research framework for this project consists of four main phases which are requirement specification, data collection, system design and result analysis.

3.1 Requirement Specification

In this phase, it is determined that opinion mining is performed manually by observing and selecting a Facebook post before the comments are extracted automatically. The post shall not contain any images since only English and Malay texts are considered for sentiment classification. Two emotion libraries (happy and unhappy) are created as the requirement to check whether the word can be categorized as having emotion (happy and unhappy) or emotionless. In addition, the Facebook post identification (ID) is needed to allow the extraction of all the comments from the selected Facebook post.

3.2 Data Collection

As a case study, data is taken from *Bahagian Pengurusan Kewangan Pelajar UiTM* Facebook page. Once the post is obtained, its comments are extracted to perform the quantification of sentiments toward a particular issue. Fig 1 shows a Facebook post with the students' responses about issues on the food allowance.

Fig 1: A Facebook post with students' responses

ISBN: 978-960-474-368-1

Students' responses (comments) on the chosen Facebook post are extracted automatically with the help of JavaScript Object Notation (JSON) library.

3.3 System Design

In the system design phase, the server, program and database are considered. Firstly, the web server is needed for data pre-processing in the comments extraction. Then, related words are collected and stored into a database for sentiment analysis. Words in the database are clustered into emotions by considering the abbreviations and numbers. Hence, the design involves several components as shown in Fig 2.

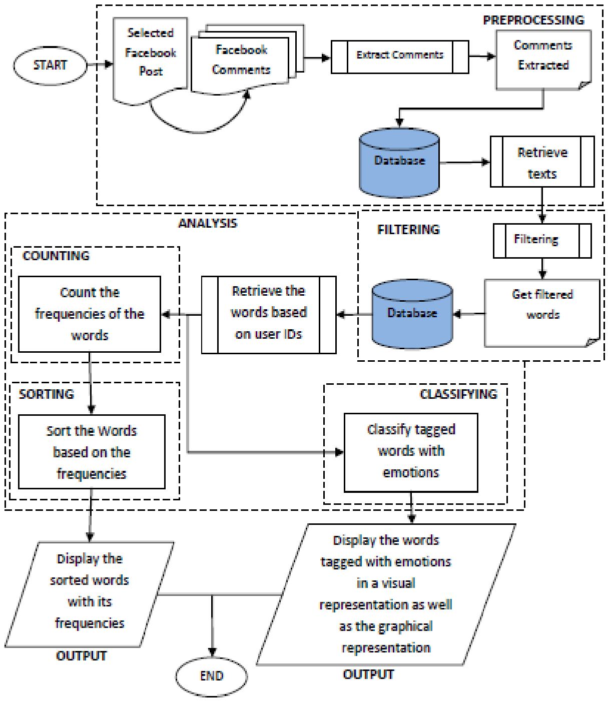

Fig 2: System Design Overview

The work flow starts with pre-processing where the comments of the chosen Facebook post are extracted and followed by filtering. After the filtering process, analysis is performed by retrieving the filtered data from the database. The analysis involves counting, sorting and classifying words which are tagged with emotions. Finally, the output is displayed to show the results.

During pre-processing, the post is chosen based on discussion issues. Then, the comments from the selected Facebook post are extracted automatically by using JSON library according to the Facebook post ID. The extracted comments are then stored in the database in order to be retrieved and filtered automatically. During filtration, the texts which consist of unwanted tags are cleaned to obtain the words including abbreviations. After the filtering process, the words that include abbreviations are stored in the database.

For the counting process, words are retrieved from the database to be counted based on the user ID. Furthermore, word frequencies are determined in order to sort the words in descending order (from the most frequently used word to the least frequently used word). Therefore, the sentiments toward the issue can be acknowledged by doing classification using the Lexicon-based approach. By doing the classifications, the words are tagged with the emotion categories. The percentages of each sentiment category are calculated. The satisfaction of the users toward a particular issue is determined by comparing the percentages between happy and unhappy sentiments. If the percentage of happy sentiment is higher than the unhappy, then it can be concluded that users are satisfied (happy) with the issue vice versa.

4 Result Analysis

The system implementation is performed by using JAVA programming language and several text files are created to handle extracted words. The text extraction is done by using JavaScript Object Notation (JSON) library to assist in developing the algorithm. Furthermore, Facebook Query Language (FQL) is used in the JSON library to get the user ID and comments.

The amount of words to be counted and sorted depends on the amount of words used by each of the users who give comments on the chosen Facebook post. The calculations are defined as follows:

Definition 1: There exists a Facebook post, P that has many users, $u_i = u_1, u_2, \ldots, u_m$.

Definition 2: For every user, u_i, there are many comments, $c_j = c_1, c_2, \ldots, c_n$.

Definition 3: For all comments, c_j, there are many words, $w_k = w_1, w_2, \ldots, w_p$.

Definition 4: For every word, w_k, the frequency is obtained where k is any arbitrary integer such that k = 1, 2,....,s. For example, word1 = w_1 and word2 = w_2.

ISBN: 978-960-474-368-1

The calculations are done by firstly retrieving a Facebook post, P. Then, from the particular Facebook post, all the users' IDs are obtained and for every user, u_i all the comments of the particular user are obtained in order to count every word used by the user. This counting and sorting of the words are done based on the selected user ID. According to the definitions as well as the descriptions, the calculations can be denoted in mathematical notations.

P: The comments of all the users commented on a chosen Facebook post are retrieved and each of the user's comments contains words. This is shown in mathematical notation in (1).

$$\mathbf{P:}\ \sum_{i=1}^{m}\sum_{j=1}^{n} u_i\ c_j,\ \text{where } w_k \subset c_j. \qquad (1)$$

From (1), each of the words obtained from the user's comments are counted to observe the frequencies of each word used by the user as shown in (2).

$$\forall \mathbf{u_i}: \sum_{k=1}^{p} w_k\ ,\ \text{where } \sum w_1\ , \sum w_2, \dots, \sum w_p. \qquad (2)$$

From (2), the words are then sorted in descending order (from the most used word to the least used word). The purpose is to view the user's most serious problem or issue quicker by looking at the most used word. After the frequencies of each word occurrences are obtained and sorted, each word is then mapped with three different colours, $r_t = \{r_1, r_2, r_3\}$ that represent sentiments. Hence, for every word, w_k, it is mapped or tagged with red, blue or gray. The classification of words with the sentiment categories that are represented in a form of colours can be denoted as follows:

r_1 = red (unhappy emotion),
r_2 = blue (happy emotion),
r_3 = gray (emotionless)

$$\therefore \forall w_k \rightarrow r_t,\ \text{where } t \text{ is either 1, 2 or 3.} \qquad (3)$$

An example of $\forall w_k \rightarrow r_t$ is as follows:

w_1 = "disappointed", therefore $r_t = r_1$ which is unhappy emotion.
w_2 = "like", therefore $r_t = r_2$ which is happy emotion.

From (3), the emotion patterns of each user for the selected Facebook post can be observed and analyzed for the sentiment analysis. The classification of every word is carried out in the classifying part.

The Facebook post identification (ID) of the Facebook page can be obtained from the URL bar and each Facebook post have different post ID. Fig 3 shows the ID location in a Facebook.

Fig 3: Facebook post ID

The post ID for each Facebook post is unique. Fig 4 illustrates the example of user ID and comments produced by the particular person.

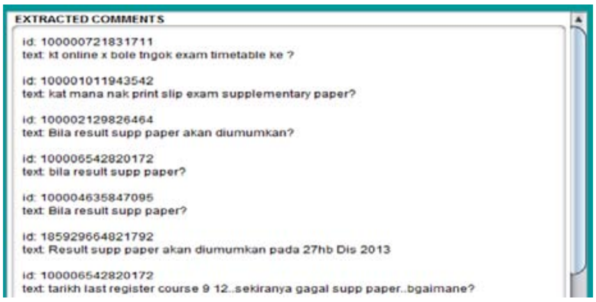
EXTRACTED COMMENTS
id: 100000721831711
text: kt online x bole tngok exam timetable ke ?
id: 100001011943542
text: kat mana nak print slip exam supplementary paper?
id: 100002129826464
text: Bila result supp paper akan diumumkan?
id: 100006542820172
text: bila result supp paper?
id: 100004635847095
text: Bila result supp paper?
id: 185929664821792
text: Result supp paper akan diumumkan pada 27hb Dis 2013
id: 100006542820172
text: tarikh last register course 9 12..sekiranya gagal supp paper..bgaimane?

Fig 4: Cleaned output of the extracted comments

During quantification process, words are classified into emotions based on two emotion libraries (text files) created earlier. For every happy words matched, the counter is increased accordingly. Results are illustrated as the percentage of each sentiment categories as shown in Fig 5.

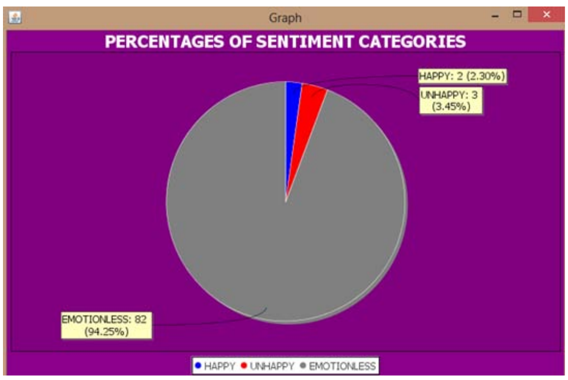

Fig 5: Percentages of the Sentiment Categories

The results show that emotionless has the highest percentage which is about 94.3% since the issues in the *Bahagian Pengurusan Kewangan Pelajar UiTM* Facebook page are mostly facts. Within the facts, the users also use happy emotion words as well as unhappy emotion words. The percentage difference between the two emotion categories (happy and unhappy) is the main consideration to quantify students' sentiments. By comparing the percentages of happy and unhappy emotion, the users' satisfaction of the issue can be determined. In order to determine the accuracy of the quantification, several different Facebook posts are chosen to be tested.

5 Conclusion

In this paper, sentiment analysis on users' emotions towards a discussion issue is presented. A computerized tool is implemented to extract words in a Facebook post and its comments so that people's emotion towards the issue can be determined if the overall feedback of people is happy, unhappy or emotionless. The results are beneficial in many cases especially in business matters when the provided service can be improved and able to attract more customers.

Acknowledgement

The authors would like to thank Universiti Teknologi MARA (UiTM) for the financial support: 600-RMI/DANA 5/3/PSI (51/2013).

References:

[1] Bargh, J. A., & McKenna, K. Y. A., The Internet and Social Life, *Annual Review of Psychology*, Vol.55, 2004, pp.573 – 590.

[2] Potena, D., & Diamantini, C., Mining Opinions on the Basis of Their Affectivity, *Collaborative Technologies and Systems (CTS), 2010 International Symposium on IEEE*, 2010, pp. 245 – 254.

[3] Bulik, B. S., Apple, Facebook tops for college students, *Advertising Age*, Vol.78, 2007, pp.12.

[4] Ellison, N. B., Steinfield, C., & Lampe, C.,The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites, *Journal of Computer-Mediated Communication*, Vol.12, 2007, pp. 1143 – 1168.

[5] Selwyn, N., Faceworking: Exploring Students' Education-related Use of Facebook, *Learning, Media and Technology*, Vol.34, No.2, 2009, pp. 157 – 174.

[6] Zhang, K., Cheng, Y., Xie, Y., Honbo, D., Agrawal, A., Palsetia, D., Lee, K., Liao, W., & Choudhary, A., SES: Sentiment Elicitation System for Social Media Data, *IEEE International Conference on Data Mining Workshops*, 2011, pp. 129 – 136.

[7] Abidin, S.Z.Z., Omar, N., Radzi, M.H.M., & Haron, M.B.C., Quantifying Text-based Public's Emotion and Discussion Issues in Online Forum, *International Journal on New Computer Architectures and Their Applications*, Vol.1, No.2, 2011, pp. 428 – 436.

[8] Akaichi, J., Social Networks' Facebook' Statutes Updates Mining for Sentiment Classification [PDF document], Retrieved from http://iir2013.isti.cnr.it/wp-content/uploads/2013/oi/iir2013_submission_24.pdf.

[9] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., & Perera, A., Opinion Mining and Sentiment Analysis on a Twitter Data Stream, *The International Conference on Advances in ICT for Emerging Regions*, 2012, pp. 182 – 188.

[10] Huang, A.H., Yen, D.C., & Zhang, X., Exploring the Potential Effects of Emoticons, *Information & Management*, Vol.45, 2008, pp. 466 – 473.

[11] Pak, A. & Paroubek, P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC), 2010, pp. 1320 - 1326.

[12] Li, J. & Ren, F., Creating a Chinese Emotion Lexicon Based on Corpus Ren-CECps, *Proceedings of IEEE on Cloud Computing and Intelligence Systems (CCIS)*, 2011, pp. 80 - 84.

[13] Lu, B. & Tsou, B.K., Combining a Large Sentiment Lexicon and Machine Learning for Subjectivity Classification, *Proceedings of the Ninth International Conference on Machine Learning and Cybernetics*, 2010, pp. 3311 - 3316.

[14] Pempek, T.A., Yermolayeva, Y.A., & Calvert, S.L., College Students' Social Networking Experiences on Facebook. *Journal of Applied Development Psychology*, Vol.30, 2009, pp. 227 – 238.

[15] Stapa, S.H., & Shaari, A.H., Understanding Online Communicative Language Features in Social Networking Environment. *Journal of Language Studies*, Vol.12, No.3, 2012, pp. 817 – 830.