

WEB MINING IN CLASSIFYING YOUTH EMOTIONS

Zura Izlita Razak¹, Shuzlina Abdul-Rahman²,
Sofianita Mutualib³ and Nurzeatul Hamimah Abdul Hamid⁴

^{2,3,4}*Research Initiative Group of Intelligent Systems,*
^{1,2,3,4}*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,*
40450 Shah Alam, Selangor, Malaysia
¹*zuraizlita94@gmail.com*, ²*shuzlina@tmsk.uitm.edu.my*,
³*sofi@tmsk.uitm.edu.my*, ⁴*nurzeatul@tmsk.uitm.edu.my*

ABSTRACT

Social media sites are websites used as mediums to create and share various types of contents over the internet. These sites can also be accessed through applications on mobile gadgets. Different social media sites are available for free, and most teenagers or youths have at least one active account. They use social media sites to connect and share their online profiles, daily activities, stories, and emotions. Depending on their social settings, their activities may or may not be seen by others. One of the latest trends that is spreading over the social media is the Korean Pop entertainment or popularly known as KPop. Over the social media, youths share and express how they feel about their Korean celebrities, music, and drama. However, the issue of excessive sharing of emotion-sharing over social media may increase the risk of mental illness and affect their mental health. Their obsession to keep up-to-date with their idols might lead or cause adverse consequences on their emotional states of mind. Thus, the aim of this research is to study the changes of youths' emotions in two different countries which are Malaysia and Korea that are related to the KPop trend. We extract texts from tweets from Twitter social media sites using the Twitter API as the basis of our study. Then, the keyword 'KPop' is used to filter the tweets. Web mining model classifies the 12,000 tweets into six emotion categories, which are joy, sadness, fear, anger, disgust, and surprise. The system then records the emotion changes and the triggering events respectively.

Keywords: Emotion analysis, KPop, Social Media, Sentiment Analysis, Twitter, Web Mining

1. Introduction

Web Mining techniques extract web documents in the form of texts, images, and audio files to explore and analyze the patterns. Web mining also can be used to extract social media users' opinions and thoughts to analyze their sentiments. Users of social media share their lives, opinions and discussions on prevailing issues via micro-blogging platforms. Some examples of current popular micro-blogging are Twitter, Facebook, and Instagram. Others include WhatsApp, Telegram, WeChat etc.

In recent years, one of the most popular topics trending on social media is the Korean Pop or KPop. In the late 1990's, KPop craze spread across Asian countries including Malaysia. The number of viewers who tuned in to YouTube channel from 235 countries in 2011 has reached up to 2.3 billion including 289,639,969 viewers among Malaysian (Seo, 2012). The Internet provides an influential social media platform that helped spread Korean trends around the world. In comparison, Korean fans tend to use Twitter to share information and emotions toward their favourite celebrities' due to its rapid speed in uploading information on the web.

Every person is free to express and share his/her thoughts on just about anything in social media. In social media, borderless access allows users to post on anything that they wish to share. On most occasions, it can cause clashes of opinions and even worse, may trigger heated debates over an issue on social media. Immature users, the youths, for example, may experience emotional changes due to nature of these issues. The changes of emotions on social media such as Twitter are reflected by the status updates (tweets) and subsequent replies (retweets). Consequently, in extreme cases, they may risk suffering from mental illnesses such as

loneliness and depressions (Pantic, 2014). This paper presents the development of the web mining application to analyze youths' emotions towards KPop in Malaysia and Korea. In Section 2, we describe the related works on sentiment analysis and in Section 3, the methodology of this work is presented. Section 4 discusses the results and findings. We conclude the paper in Section 5.

2. Sentiment Analysis

Sentiment analysis is the process of determining the opinions, attitudes, evaluations, emotions, and reviews expressed in texts towards any aspect of businesses such as products or brands or a public opinion behind certain topics. These opinions are usually classified as positive, negative or neutral. Applying sentiment analysis helps the organisation understand their customer better and be more proactive about the changing dynamics in the market place. For example, an opinion can be extracted from an organization's internal data, which are usually customers' feedbacks from emails.

Opinions can also be extracted from news articles and word-of-mouth commentaries on the web such as individual experiences, opinions, comments on articles or issues, and postings on social networking sites. Examples of the applications of the analysis include attempts by businesses and organizations to seek customers' opinions using consultants and surveys, on what are the considerations used to decide to purchase products or use services or get public opinions about political issues, advertisement placements to place an advertisement if people like the products and opinions retrieval to provide general search for opinions.

The popularity of micro-blogging is increasing as a communication platform on the web. It allows users to broadcast their opinions or thoughts to the public. The texts that broadcasted via Twitter, a micro-blog, is known as tweets. Every tweet has a maximum of 140 characters in length, thus allowing the users to get and propagate information effortlessly (Yoo et al., 2018). Users of Twitter can broadcast several types of information such as conversations, comments on issues, news reporting and updates on current events.

Researchers tend to use sentiment analysis on social media applications to analyze various issues. Among the many popular techniques in sentiment analysis include Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (Birjali et al., 2017; Narayanan et al., 2013; Neethu & Rajasree, 2013). Birjali et al., (2017) for example, compared several machine learning algorithms techniques to predict suicide sentiments using Twitter data. Narayanan et al., (2013) studied the attitude of a speaker or a writer with respect to some topics or simply the contextual polarity of a document. In that study, Narayanan et al., (2013) employed Bernoulli NB with the enhancement of Laplacian smoothing and handling negations was used. Meanwhile, Neethu & Rajasree (2013) proposed SVM and NB to analyze Twitter posts on electronic products. NB Classifier makes use of all the features in the feature vector and analyzes them individually as they are equally independent of each other. The accuracy of NB algorithm in the research was at 89.5%.

Qamar & Ahmad (2015) proposed detection on emotional content from texts in which the emotions are categorized into six types, which are happy, surprise, sadness, fear, anger and disgust. From the six categories of emotions, the emotions are then classified into two, which are positive and negative emotions. In another study, Nandhini & Sheeba (2015) identified the presence of cyber bullying terms and classifies cyber bullying activities in social network into types of behaviors such as flaming, harassment, racism and terrorism. All these studies have shown that sentiment analysis is beneficial in analyzing user's attitudes.

3. Methodology

The social media mining in web architecture is developed with the aim of classifying six types of emotion categories namely, joy, sadness, fear, anger, disgust and surprise. In this study, the texts are extracted from Twitter API. The focus of the extraction is on youths' interests towards Korean entertainment and the keyword used is 'KPop' and has been extracted from

two different countries which are Malaysia and Korea. The number of data tweets collected is 3000 tweets each.

The architecture of this system comprises of five components. Figure 1 shows each of the components of this architecture. The first component is data pre-processing. The data pre-processing component consists of data cleaning, normalization and tokenization processes. In data cleaning, the function and empty words are removed. Then, normalization process converts all the words from upper case to lower case. Finally, the tokenization process removes the numbers and symbols from the data. In addition, the html link is removed, other than that, #, mention (people), punctuation, numbers, unnecessary spaces, NA value and repetitive tweets are also dropped.

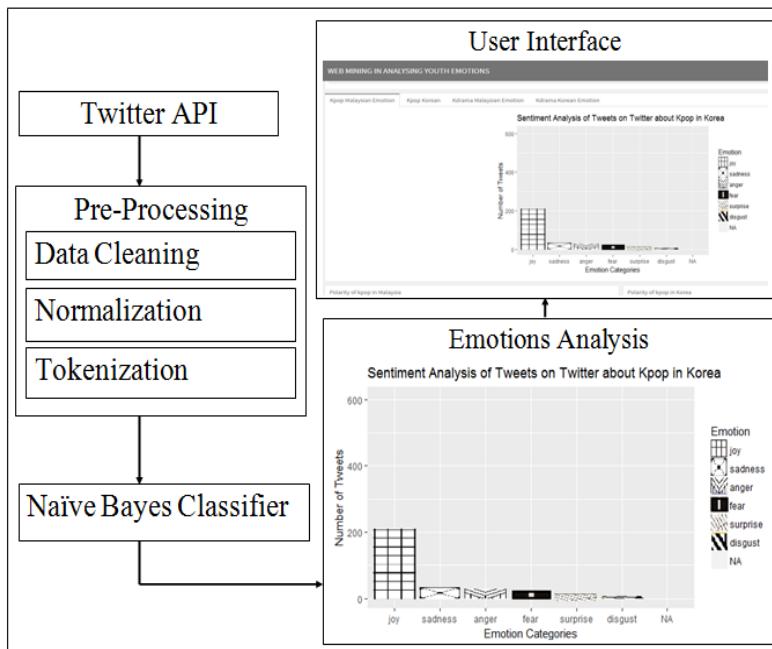


Figure 1 System Architecture

The pre-processed tweets are then classified into six emotions categories in the second component of the system, which is the Naïve Bayes Classifier (NBC). The NBC model is built by training the dataset of collected tweets with corpus of emotion words. The classifications of the emotions are based on the standard emotion corpus. The standard emotion corpus contains 14,182 words associated with the six emotions categories as mentioned earlier.

The third component is to visualize the results by using a histogram graph. The histogram illustrates the number of tweets corresponding to each emotion category. The last component is the system interface. The interface is designed to appear as dashboard. The dashboard displays all the graphs of Malaysian and Korean youths' emotions based on the KPop keyword through html Web page.

We developed NBC as a classifier for emotions from tweets. The extracted tweets are then partitioned into a single word. Every word is linked to the emotion corpus. Figure 2 shows an example of tweets that contains ten words. Stop words are filtered out after the pre-processing of the natural language data. Stop words refer to the most common words in a language. For instance, the stop words of the tweet in Figure 2 are “a”, “the”, and “guys” are removed in Figure 3. These words are then replaced with the available word in the corpus as shown in Figure 4.

give	a	listen	and	a	Like	these	guys	are	Awesome
------	---	--------	-----	---	------	-------	------	-----	---------

Figure 2 Example of tweets

give		listen	and		Like				Awesome
------	--	--------	-----	--	------	--	--	--	---------

Figure 3 Example of tweets after removing the stop words

Neutral		Neutral	Neutral		JOY				Neutral
---------	--	---------	---------	--	-----	--	--	--	---------

Figure 4 Example of tweets after replacing with the corpus

In Equation 1 (Jurafsky & Martin, 2017), the probability of the emotion category is calculated in every word. The categories retrieved are then compared. The emotion category with the highest probability value is chosen to be the best fit.

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c|d) = \operatorname{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

For a sentence referred to as document d , out of all classes $c \in C$, the classifier returns the class \hat{c} , which has the maximum posterior probability given the document. Bayes rule in Equation 1 is computing $\frac{P(d|c)P(c)}{P(d)}$ for each possible class. $P(d)$ does not change for each class, which must have the same probability $P(d)$. Equation 2 computes the most probable class \hat{C} , given some documents d by choosing the class, which has the highest product of two probabilities, which are the prior probability of class $p(C)$ and likelihood. Equation 2 is extended into Equation 3 (Jurafsky & Martin, 2017).

$$\hat{c} = \operatorname{argmax}_{c \in C} P(c|d) = \operatorname{argmax}_{c \in C} P(d|c)P(c) \quad (2)$$

$$\hat{c} = \operatorname{argmax}_{c \in C} \underbrace{P(f_1, f_2, \dots, f_n|c)}_{\text{likelihood}} \underbrace{P(c)}_{\text{prior}} \quad (3)$$

Table 1 shows the probability of the emotions of each category and the last column is the best fit. For a KPop in Korea, the example of the tweets classification after using the stop words are shown in Figure 5. Table 2 states the *best_fit* emotion after Naïve Bayes algorithm was applied. The first row is the probability of the tweets.

Table 1 Emotion classification of KPop tweets

#	Anger	Disgust	Fear	Joy	Sadness	Surprise	Best_Fit
1	1.50	3.09	2.07	1.03	1.73	2.79	NA
2	7.34	3.09	2.07	1.03	1.73	2.79	Anger
3	1.47	3.09	2.07	1.03	1.73	2.79	Na
4	1.47	3.09	2.07	7.34	1.73	2.79	Joy
5	1.47	3.09	2.07	7.34	1.73	7.34	Joy

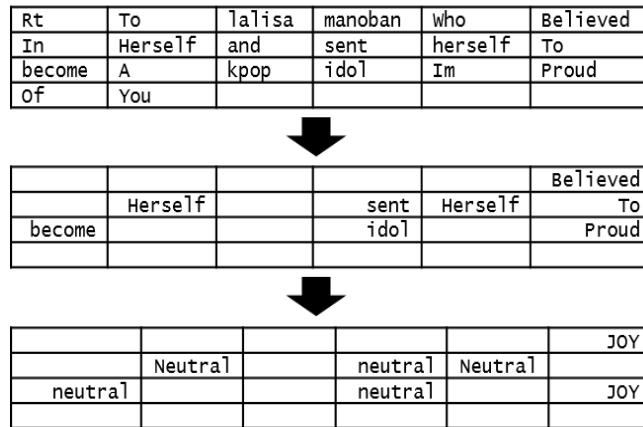


Figure 5 The process of stop words filtration and emotion classification based on standard e motion corpus

Table 2 Korean KPop tweets emotion classification

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Best_Fit
1.	1.50	3.09	2.07	7.34	1.72	2.79	Joy
2.	1.50	3.09	2.07	1.05	1.72	2.70	NA
3.	1.50	3.09	2.07	7.34	1.72	2.79	Joy

4. Result and Discussion

The emotions extracted from of the KPop tweets from Malaysian and Korean youths were then used to plot the graph. The analysis of the tweets was done by extracting 6,000 tweets. The number of tweets to be analyzed decreased after the pre-processing. Consequently, the number of users' emotions based on the six types of emotion also decreased due to the records. Therefore, the tweets with the neutral emotion are saved as 'NA'. Figure 6 is the histogram graph based on KPop in Malaysia tweets. The highest emotion of the group of people who updated about KPop is 'Joy' with tweets of more than 250. The second highest emotion is 'Sad' with 50 tweets and followed by 'Anger' with 45 tweets, 'Surprise' with 30 tweets, 'Fear' with 15 and 'Disgust' with the lowest number of tweets at 10. Figure 7 is the histogram graph based on Korean KPop tweets. The highest emotion of youths who updated about KPop is 'Joy' with more than 250 tweets. The second highest emotion is 'Sad' with 50 tweets followed by 'Anger' with 48 tweets.

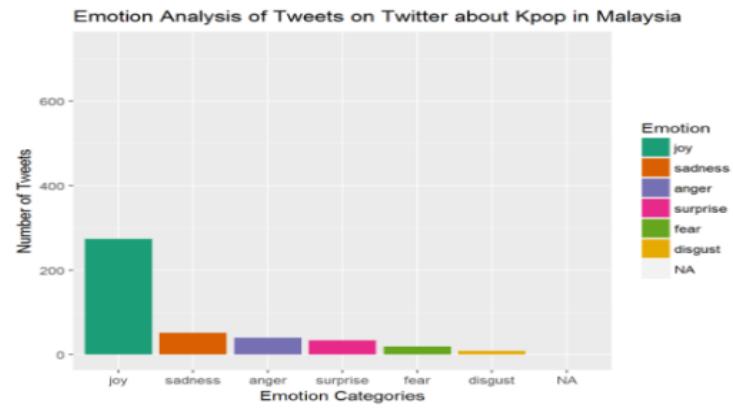


Figure 6 Visualization on KPop emotion in Malaysia

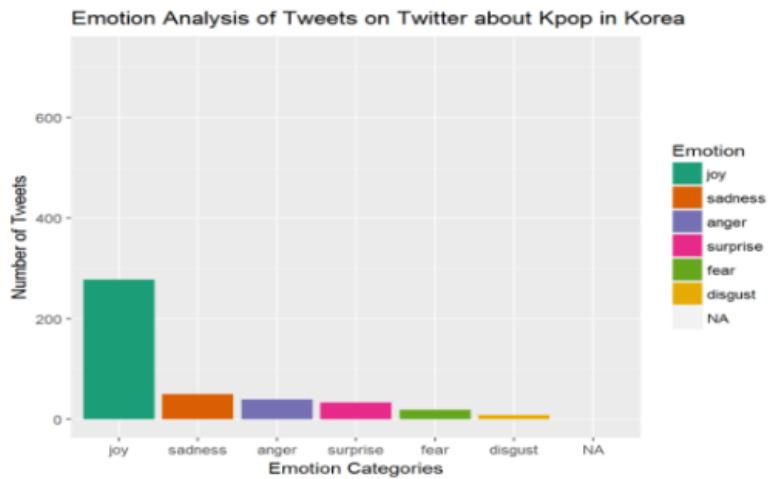
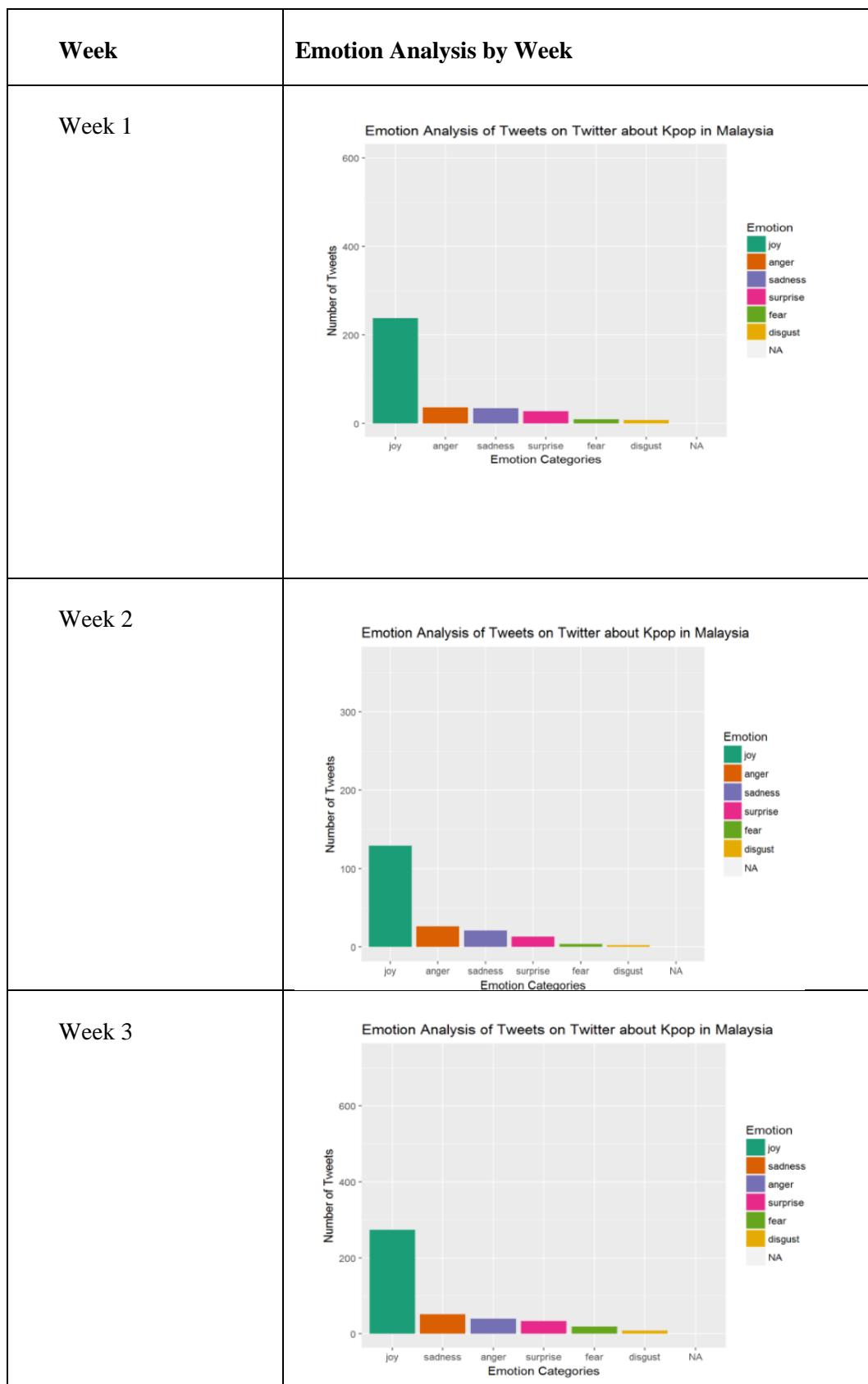


Figure 7 Visualization on KPop emotion in Korea

The comparison of the emotions was also made in July 2017. The comparison was done based on 3 weeks of emotion records within one month. In July 2017, the stages of emotions from KPop tweets in Malaysia for three weeks, it was observed that the top emotion was the same which is ‘Joy’. For the second emotion, on 8th and 15th July, the emotion was anger. On 22nd July, the emotion changed to ‘Sad’. For the third emotion, on 8th July and 15th July, the emotion was ‘Sad’, and on 22nd July, the emotion changed to ‘Anger’. For the following emotion, the emotion states remained the same, which was the emotions of ‘Surprise’, ‘Fear’ and ‘Disgust’. Table 3 shows the emotion changes in Malaysia towards KPop in July 2017.

Table 3 Emotion analysis of KPop in Malaysia in July 2017



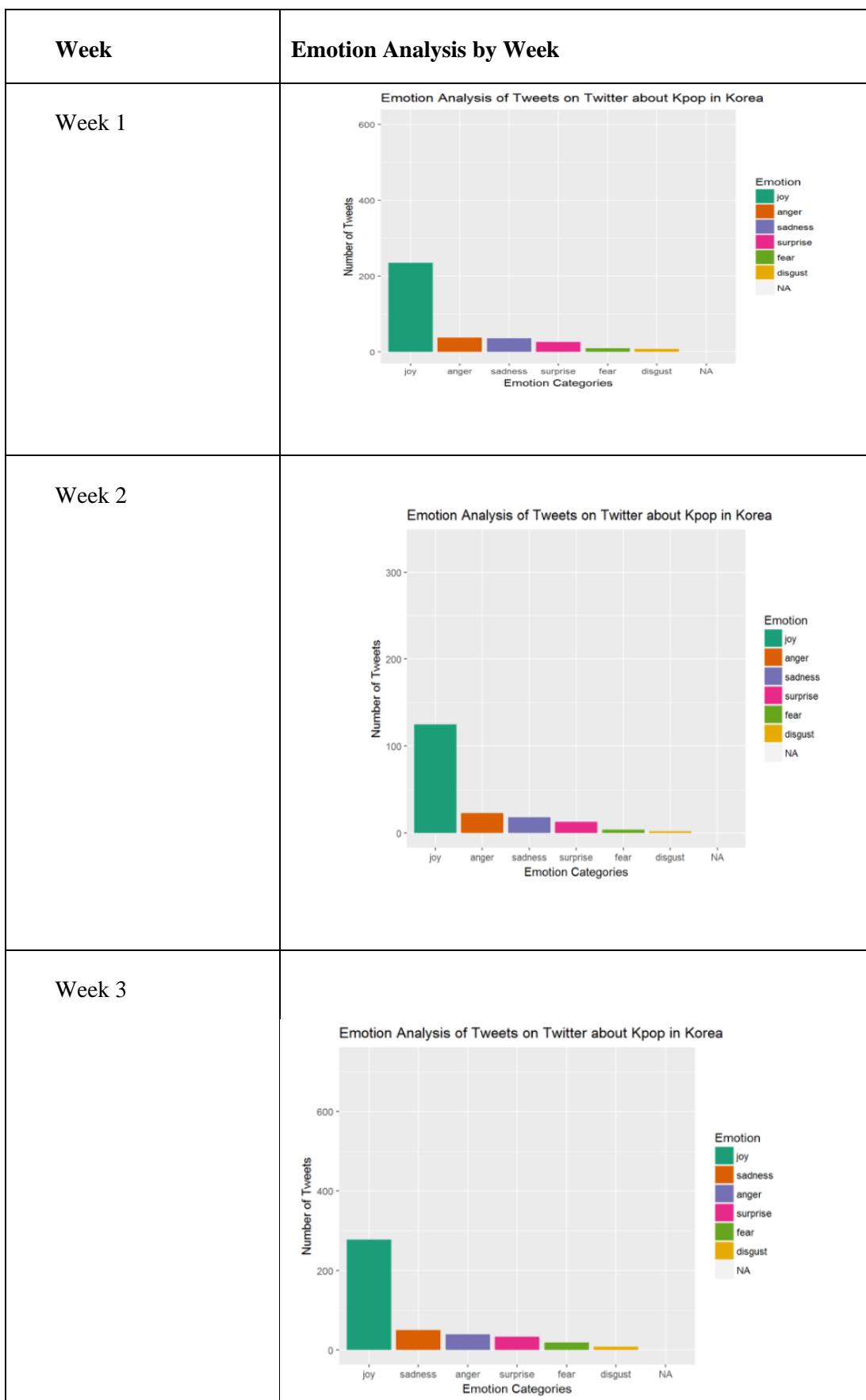
The series of triggering events that changed the youths' emotions in Malaysia towards KPop in July 2017 is illustrated in Figure 8. For 'Joy' emotion, the most frequent emotion word that appeared was '*like*'. During this period, the Korean idol comebacks were the NU'EST W, Black Pink and K.A.R.D. For the 'Sad' emotion, one of the events which occurred was one of reality shows to search for boy groups with 101 participants ended. 11 of them were chosen to debut using the group name 'Wanna One'. The emotion of 'Anger' was associated with an event featuring the Korean idol, EXO who came out with their new album. The fans and haters of the EXO group had arguments on social media. For the 'Surprise' emotion, the most frequent word was '*top*'. During this period, the Korean singer, Choi Seung-Hyun with the stage name TOP was arrested for smoking marijuana.



Figure 8 Malaysian KPop word cloud

Table 4 shows the emotion changes among Korean youths towards KPop in July 2017. As can be seen, the most dominant emotion among Korean youths' was 'Joy'. Between the first and second week, the second dominant emotion was 'Anger'. However, in the third week, it changed to 'Sad'. Meanwhile, the third dominant emotion in the first and the second week was 'Sad'. But it changed to 'Anger' in the third week. For the other emotion states of 'Surprise', 'Fear', and 'Disgust', they remained the same.

Table 4 Emotion analysis of KPop in Korea in July 2017



There were a series of triggering events of youths' emotions in Korea towards KPop which occurred in July 2017. Figure 9 illustrates the word cloud that highlights the events. The emotion of Korean and Malaysian Twitter users towards KPop were similar for several events. For example, the emotion '*Surprise*' was similar when TOP was arrested for smoking marijuana. The '*Joy*' emotion also associated with KPop idol comeback and several KPop stars birthday wishes.



Figure 9 Korea KPop word cloud

5. Conclusion and Future Works

This research analyzed the emotion changes between youths in Malaysia and Korea towards KPop. This research analyzed the emotion changes within several weeks and months according to the different triggering events on Twitter. By using NB algorithm, the model can classify the tweets according to the maximum posterior probability which is suitable for natural language text. Consequently, it can be extended by deploying the corpus that using Malay and Korean languages to analyze youths' emotions. Moreover, the charts produced would help us to understand the patterns of tweets over interested period, together with the common words occur. Another possible extension of this research is to extract social media posts from other platforms such Instagram and Facebook. The analysis from this research is also beneficial for external features such as notifications or alarming systems.

Acknowledgement

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia for the support throughout this research.

References

- Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. *Procedia Computer Science*, 113, 65-72.
- Jurafsky, D., & Martin, J. H (2017). Speech and Language Processing. All rights reserved. Draft of August 7, 2017. Copyright c 2016. All rights reserved.
- Nandhini, B. S., & Sheeba, J. (2015). *Online social network bullying detection using intelligence techniques*. Procedia Computer Science, 45, 485-492.
- Narayanan, V., Arora, I., & Bhatia, A. (2013). *Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model*. In H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee,
- Neethu, M., & Rajasree, R. (2013). *Sentiment analysis in twitter using machine learning techniques*. Paper presented at the Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013.
- Pantic, I. (2014). *Online Social Networking and Mental Health. Cyberpsychology, Behavior and Social Networking*, 17(10), 652–657. <http://doi.org/10.1089/cyber.2014.0070>
- Qamar, S., & Ahmad, P. (2015). *Emotion Detection from Text using Fuzzy Logic*. International Journal of Computer Applications, 121(3). Pg no
- Seo, M., S. (2012). *Lessons from K-pop's Global Success*, 5(3), 60. Retrieved from <http://connection.ebscohost.com/c/articles/77971026/lessons-from-k-pops-global-success>
- Weise, T., Li, B., & Yao, X. (Eds.) (2013). Intelligent Data Engineering and Automated Learning – IDEAL 2013: 14th International Conference, IDEAL 2013, Hefei, China, October 20-23, 2013. Proceedings (pp. 194-201). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Yoo, S., Song, J., & Jeong, O. (2018). Social media contents-based sentiment analysis and prediction system. *Expert Systems with Applications*, 105, 102-111.