Word-forest Visualization of Discussed Topics in Social Media Comments

M. Bakri[1,2], SZZ. Abidin[2], N. Omar[2], M. Hamiz[1] and Afiq Razali[1]
[1]*Universiti Teknologi MARA Cawangan Melaka (Kampus Jasin),*
[2]*Universiti Teknologi MARA Malaysia.*
bakri@tmsk.uitm.edu.my

***Abstract*—It becomes a norm for many organizations to use social network as a platform for internal and external communication means. Due to its extensive usage, most large organizations recognize the importance of capturing disseminated information across the social networks for the benefit of their internal perusal. However, managing and keeping track of all the information which are hidden in the piles of comments are hard to deal with. This paper presents a system that can extract, analyze and visualize information from the comments. As for the case study, Facebook is chosen due to its ability to allow people to comment freely and repetitively. The comments were extracted from selected post in Facebook using its API. The relationship between the words inside the comments will then be determined by using relationship table. Then, a visualization technique, word-forest, is used to visualize the relation between the prepared table. The prototype is tested by using selected posts in specific Facebook accounts. The result shows that users can quickly get overviews on the topics that have been discussed without having to go through all the comments on the Facebook. The system has great potential to be further explored as one of the means to get internal and external workers or public perception unobtrusively at real-time and real-life setting.**

***Index Terms*—Data Mining; Data Visualization; Real-time Information; Social Network.**

I. INTRODUCTION

In this digital world, social media platform has become a popular medium for people to interact with their relatives or friends and exchange information among them. In addition, it also acts as a platform for people to voice out their opinions or thoughts that cross their mind. Through the platform such as Facebook and Twitter, feelings or opinions can be shared interactively with other users in a matter of seconds. The exchanged information and opinions offer huge amount of real-time data which can be analysed to allow a better understanding of dominant trends and patterns, that can be used for decision support (e.g., design a better product to fulfill the users' needs, making more effective marketing campaigns) [1]. However, this data is usually enormous and naturally noisy. To obtain a full picture of a discussion by sifting through individual comments manually is very challenging if not impossible [2].

Even though all the comments or responses able to be collected easily, interpret all the opinions is the real challenge. Popular posts in Facebook for example usually contains thousands of comments. This huge amount of data makes it difficult for the owner of the post to uncover and extract any helpful and meaningful information from them.

II. LITERATURE REVIEW

This section highlights related research in the area of social media and related approach in data visualization.

A. Social Network Data Analytics

Social media platform allows the public to discuss ideas or sharing opinions about certain issues related to a post using comments. This kind of data may able help researchers to identify what is currently trending. Government and businesses, for example, can collect and analyze this data for a certain purpose in the future. For example, they started to be involved in this kind of social network analysis as their strategy to improve their service [3].

B. Related Work

Some existing works on text summarization for social media platform has been done. In [3], a method was proposed to produce journalistic summaries for a sport event by extracting related Twitter status updates. Temporal cues, such as spike in the number of status updates during the event are used to identify important moments. Each important moments within the event are described by using a phrase graph which was extracted from the corpus of status updates the longest relevant sentences. In [4], a topic summarization framework was proposed to extract the temporal correlation that exists among tweets by using Decay Topic Models. The framework effectively extracts meaningful topics which capture different aspects of the sport event. Other related works in social network text summarization are [5]–[7]. However, most of the works are using algorithm alone to do the summarization. Hence, this work proposes data visualization technique to aid the summarization process.

C. Data Visualization

Visualization enhances the presentation of complex data, assisting human cognition, and allow for a user to investigate large amount of data in one setting [8], [9]. Relationships and patterns that may go unnoticed in content-based information can be discovered and perceived easier using data visualization [10].

To build an effective visualization, it must be able to precisely and proficiently deliver the desired information to the user. The goal of visualization is to translate raw data into visual or graphical representation which can easily, accurately, efficiently and meaningfully interpreted. The message that is to be delivered by the visualization is Quantitative Message. There are eight types of Quantitative Message which are: ranking, time-series, correlation, nominal comparison, geographic or geospatial. The objective

of this project is to extract and summarize relevant opinions from a large amount of comments. Thus, the ranking and correlation will the focus in the visualization.

Text visualization able to group the common words and helps users to see the relationship between them and discovering patterns. It able to help see the data more clearly and understanding the raw data or big data in a more convenient display. There are a few visualization techniques such word cloud and word tree.

Word cloud is used to discover words that appear more frequently in the raw data [11]. The size of the words in the visualization refers to their frequency where bigger word means that it appears more frequently in the data than the smaller one. However, word cloud cannot show the relationship between the words. Another technique that is commonly used is word tree. A word tree is essentially an interactive type of keyword in context (KWIC) system [12]. The layout of word tree visualization is shown in Figure 1.

The outline makes clear the natural tree structure of the context. It allows easy approaches to examine the context further. Due to the hierarchical structure, words in a subtree cannot have relationship with words in another subtree.

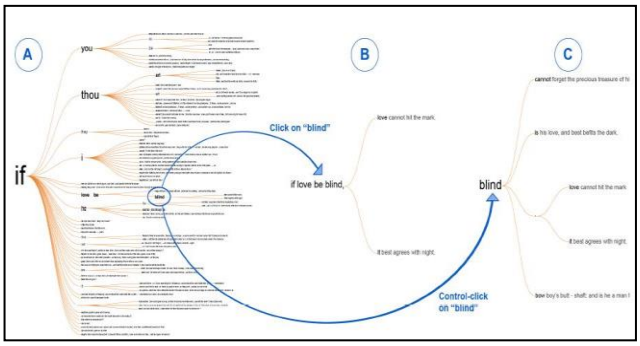

Figure 1: Word Tree Visualization

A novel visualization named word-forest is designed based on a word cloud and word tree. The different word size in word cloud enables users to spot higher frequency words based on their ranking and the edges between the words enable users to spot relationship or correlation between those words. These may help users in identifying trending topics being discussed easily.

III. METHODOLOGY

A prototype is developed to test the visualization. Development of the prototype follows a normal Software Development Life Cycle (SDLC). The prototype consists of four main modules, which are data extraction, data cleaning, word relationship and word frequency table, and data visualization module.

A. Data Extraction

This module is responsible for extracting data from Facebook comments. Facebook is chosen as the social media platform due to its large user base and large communities. To extract the data, Facebook4J, java API for Facebook will be used. Comments will be extracted from a post that has been chosen by the users. The extracted comments are stored inside a text file for processing.

B. Cleaning

Cleaning module purpose is to remove all the common words from the extracted comments. Common words are removed to reduce the number of words in the visualization, which can hamper the visualization result. List of common words is stored inside a text file that acts as the common word database. All occurrences of common words inside the extracted comments will be removed and the cleaned file is saved in a new text file. Currently, the common words database is only for Malay language.

C. Word Relationship and Word Frequency Table

The comments that have been cleaned and saved will be used to construct a word relationship table. Word relationship table is used to determine the relationship between words. It shows how many times a word appear right after another word inside the comments. Table 1 shows the table format. The value inside the table shows the frequency of how many times a word appear right after another word (e.g: fij is the frequency of word i appear right after word j).

Table 1
Word Relationship Table

	Word 1	Word 2	…	Word n
Word 1	f_{11}	f_{12}	…	f_{1n}
Word 2	f_{21}	f_{22}	…	f_{2n}
…	…	…	…	
Word n	f_{n1}	f_{n2}	…	f_{nn}

The pseudocode for the algorithm to construct the table is shown below.

```
relation[n][n] = {}

for each line in comments:
    for each unique word w1:
        for each unique word w2:
            currIdx=0
            while line.indexOf(w1, currIdx) != 0:
                currIdx=line.indexOf(w1, currIdx):
                if (line.indexOf(w2, currIdx) ==
                    currIdx + length(w1) + 1:
                    relation[w1][w2]++
```

Another table that will be constructed is the word frequency table. This module will count the occurrences of each unique word that exists inside the cleaned comments. Table 2 shows the table format.

Table 2
Word Frequency Table

	Frequency
Word 1	f_1
Word 2	f_2
…	…
Word n	f_n

The values in both tables will be used in the visualization to show the degree of relationship between words and the popular words used in the comments. Visualization is used as comprehending a large amount of data in the table manually is not practical and it is not able to show hidden relationship between the words.

D. Visualization

Information obtained from word relationship and word frequency module will then be projected in a form of

convenient word graph visualization. Our visualization technique, word-forest, which combines both word cloud and word tree visualization techniques, will project the word relationship and word frequency information. The word frequency will be represented by the word size that acts as a node in the graph. Word relationship, on the other hand, will be represented by edges between the word nodes. The width of the edges shows the degree of relationship between the words where higher degree will have wider width and vice versa. To make the visualization easy to comprehend, only 10% of highest frequency words will be displayed.

IV. RESULTS

The accuracy of the developed prototype was tested first. We run the prototype with a post with a small number of comments and verify the relationship and frequency table generated. We also make sure that the visualization generated does reflect the comments. Then, the developed prototype was tested with several selected public posts in Facebook in Malay language. The first post that was tested is a public post from Harian Metro's page with id *"10154154273267052"*. The visualization obtained is shown in Figure 2.

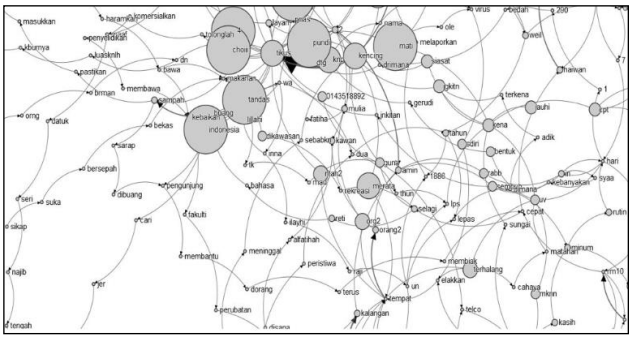

Figure 2: Visualization result for post 10154154273267052

As shown in the Figure, there are too many nodes being displayed in the visualization which makes it difficult to comprehend. A modification was made so that the visualization only shows the relationship between words where its frequency is greater than one in Word Relationship Table. This is to highlight only significance relationship between the words.

After the modification, the updated visualization is shown in Figure 3. The number of nodes being displayed was reduced significantly which allows easier interpretation of the visualization.

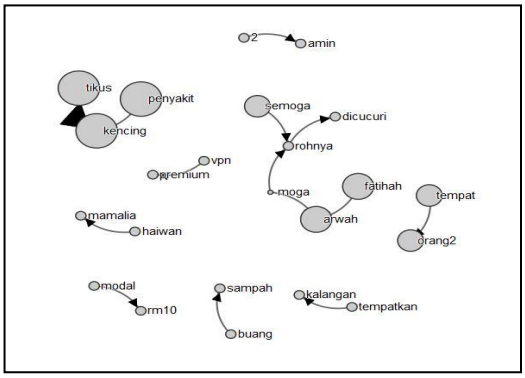

Figure 3: Visualization result for post 10154154273267052 #2

The system was then further tested with different public posts. The first post is from a news page, Astro Awani with post id *"10153598602005965"* titled *"Belum ada keputusan cukai minuman bergula"*. The post is about government suggesting on implementing tax on sugary drinks. This post contains 140 comments in total. Visualization generated by the prototype for the comments are shown in Figure 4.

From the figure, we can see that there are a few topics being discussed in the comments. The main topic is *"kesihatan rakyat malaysia konon"*, which means *"purportedly for the health of malaysians"*. Another discussed topics that can be seen are *"haramkan makanan berminyak"* which means "ban oily food" and *"tutup kilang"* which means, *"close the factory"*.

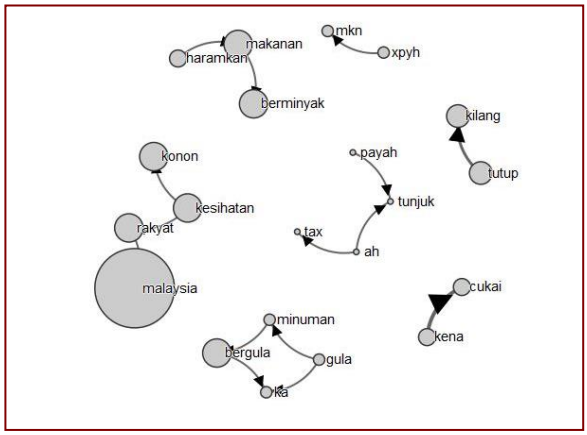

Figure 4: Visualization result for post 10153598602005965

The second post that was selected for testing is also from Astro Awani page. The post id is *"10153601039610965"* titled *"PRK Kuala Kangsar: DAP salahkan PAS atas kekalahan"*. The news is about by-election in Kuala Kangsar. The post contains 362 comments. Visualization generated by the prototype is shown in Figure 5.

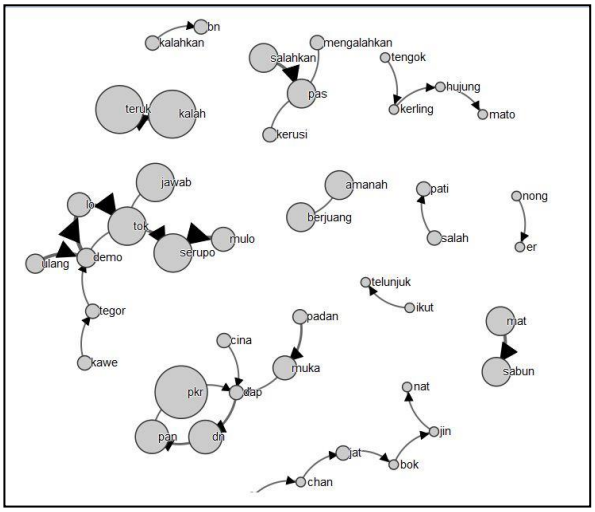

Figure 5: Visualization result for post 10153601039610965

From the figure, there are a few discussion topics that are being discussed. Some of them are *"teruk kalah"* means *"big loss"*, *"salahkan PAS"* means *"blame PAS"*, *"jawab tok serupo mulo"* means *"answer differ than previous"*.

V. CONCLUSION

The developed prototype successfully extracts, process and visualizes comments in social network Facebook. The designed visualization, word-forest, able to help users in identifying what are the topics being discussed in the comments as well as their popularity. It will able to help any parties to keep track public opinions in their social media without having to go through each comment one by one. This will help them in their decision-making process.

Currently, the prototype only supports Malay language. It can be easily used for posts in another language by swapping the stop word database. Although the prototype is tested for Facebook, it can be used for any social media comments.

For future research, the system can be further improved to handle short forms of spelling since most of the comments are using them. Ability to stem the words and group words with similar root words also may improve the visualization obtained.

ACKNOWLEDGMENT

The authors are grateful to the Research Management Centre (RMC) UiTM for providing and supporting the research (600-RMI/RAGS 5/3 (10/2014)).

REFERENCES

[1] P. Cortez, N. Oliveira, and J. P. Ferreira, "Measuring User Influence in Financial Microblogs," Proc. 6th Int. Conf. Web Intell. Min. Semant. - WIMS '16, pp. 1–10, 2016.

[2] J. Xu and T.-C. Lu, "Seeing the Big Picture from Microblogs : Harnessing the Crowdsourcing Power for Visual Event Summarization," 20th Int. Conf. Intell. User Interfaces, pp. 62–66, 2015.

[3] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," IUI '12 Proc. 2012 ACM Int. Conf. Intell. User Interfaces, pp. 189–198, 2012.

[4] S. A. Freddy Chong, Tat Chua, "Automatic Summarization of Events From Social Media," Int. AAAI Conf. Weblogs Soc. Media, pp. 81–90, 2013.

[5] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entity-centric topic-oriented opinion summarization in twitter," Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '12, p. 379, 2012.

[6] X. Yang and Y. Ruan, "A Framework for Summarizing and Analyzing Twitter Feeds," Kdd´12, no. Figure 1, pp. 370–378, 2012.

[7] W. Yin, T. Mei, and C. W. Chen, "Automatic Generation of Social Media Snippets for Mobile Browsing," pp. 927–936, 2013.

[8] S. Card, J. Mackinlay, and B. Shneiderman, "Readings in information visualization: using vision to think," p. 712, 1999.

[9] X. Lin, "Map displays for information retrieval," J. Am. Soc. Inf. Sci., vol. 48, no. 1, pp. 40–54, 1997.

[10] K. A. Stofer, "Visualizers, visualizations, and visualizees: Differences in meaning-making by scientific experts and novices from global visualizations of ocean data," p. 24, 2013.

[11] W. Cui, Y. Wu, S. Liu, F. Wei, M. Zhou, and H. Qu, "Context-preserving, dynamic word cloud visualization," IEEE Comput. Graph. Appl., vol. 30, no. 6, pp. 42–53, 2010.

[12] M. Wattenberg and F. B. Viégas, "The word tree, an interactive visual concordance," IEEE Trans. Vis. Comput. Graph., vol. 14, no. 6, pp. 1221–1228, 2008.