International Journal of Engineering & Technology, 7 (4.38) (2018) 955-959

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Visualization of Crime News Sentiment in Facebook

M. Bakri C. Haron1, Siti Z. Z. Abidin2, N. Azmina M. Zamani3

1Faculty of Computer and Mathematical Sciences,
UiTM Cawangan Melaka Kampus Jasin, Malaysia
2Advanced Analytics Engineering Centre (AAEC), FSKM,
Universiti Teknologi MARA Shah Alam, Malaysia
3Faculty of Computer and Mathematical Sciences,
UiTM Cawangan Perak Kampus Tapah, Malaysia
1bakri@tmsk.uitm.edu.my, 2zaleha@tmsk.uitm.edu.my, 3azmina@uitm.edu.my

Abstract

Facebook has become a popular platform in communicating information. People can express their opinions using texts, symbols, pictures and emoticons via Facebook posts and comments. These expressions allow sentiment analysis to be performed by collecting the data to obtain the public's opinions and emotions toward certain issues. Due to a huge amount of data obtained from Facebook, proper approaches are required to cater the texts and symbols used in the comments. There are also limited amount of dictionary on Malay texts which make it more challenging to process and classify the positive and negative words used in the comments. Thus, hybrid approach is applied during the data processing to visualize the results. In this work, a combination of lexicon-based approach and Naïve Bayes are used. This study focuses on analyzing the public's sentiments on crime news in Facebook by using word cloud visualization. The visualization displays important words used in a form of a word cloud. Moreover, the percentage of positive and negative words existed in the comments is also shown as part of the visualization results.

Keywords: *Crime news, Emotion, Sentiment classification, Social network,, Visualization..*

1. Introduction

Social media permits the interchange and establishment of user-generated content (UGC) and it is described as a set of web-based program that was established on the technological and ideological basis of Web 2.0 [1]. It is also known as a group of new media that permit social interaction between users [2] and it is undeniably a vital part in social and political developments [3]. People are allowed to spread data, opinions, statements and behavior via the social media [4]. One of the important issues is crime where public's emotions, impressions, viewpoints and judgments regarding crime can be affected by social media [5]. In addition, people tend to rely upon resources published by the media production in observing crime topics [6].

Sentiment can be classified as feeling, emotion or opinion of a person towards something such as situation or events. Emotions constitute a key element in human instinct and behaviour as it can be communicated through different media, for example, speech, facial expressions, gestures and textual information [7]. Organizations that gather and break down information from social media, news and other information streams are confronted with a few difficulties that involve storing and handling of large amount of information. Hence, it is a challenge to gather all the related information from different sources in order to analyze and present it in such a way for non-technical domain experts to understand and self-interpret the valuable significant information [8]. Therefore, a combination between sentiment analysis and interactive visualization are suggested.

This paper is divided into five Sections where Section 2 describes the preliminary study on the related work while Section 3 describes the design process in producing the visualized output by using hybrid techniques. Then, Section 4 presents the results and analysis of the system followed by Section 5 which concludes the work.

2. Preliminary Study

There are many social media on the Internet and some of them are very popular among the public such as Facebook, Twitter and Instagram. From the observations done on these three types of social media [9-12], Facebook is the most suitable platform to perform sentiment analysis due to its favourable social networking site [10] with 1.49 billion frequent users. It is also the most flexible social media as it does not have any limitations on characters' post and various types of sentiments can also be viewed on a single post. Moreover, Facebook permits public to be involved in discussions as it allows data sharing [9]. Nowadays, the way news are conveyed and events are published have been revolutionized, where Facebook has come into view as progressively important channels of interactive media content around incident and news [13]. With this, public have taken the time to express their sentiments towards events and online news. From the news in Facebook, information can be collected for analysis to understand more about the public's sentiments.

Sentiment analysis or opinion mining allows the positive, negative or neutral perspective summary on certain issues to be obtained [14]. For such analysis, machine learning is implemented to ex-

tract the textual contents for the sentiments [15]. Due to the challenge of interpreting the text [16], data visualization is used to aid in representing the information retrieved so that it can be more understandable [17]. In order to retrieve the polarity of individual words, lexicon-based sentiment analysis approach is used to categorize positive and negative sentiments using word matching [18-19]. From the lexicon-based approach, dictionary-based approach is applied instead of corpus-based approach. The corpus-based approach is lacking of standard corpuses which make it difficult for performing sentiment analysis on Malay words [20]. The analysis is based on a supervised machine learning technique since the output is meaningful to humans. Hence, Naïve Bayes is applied as the training classifier. Naïve Bayes classifier has high speed and precision, and it is one of the most famous techniques [21]. Therefore, the combination of lexicon-based approach and Naïve Bayes technique is applied in clustering Malay texts to produce a visualized output. Some of the previous related works also use hybrid approach as it provides better accuracy [22]. High accuracy in determining sentiment polarity and strong sentiment detection was achieved in hybrid approach [23]. Even though, the lexicon-based approach has high precision, it also has low recall as it depends fully on the opinion words to determine the polarity of sentiment [24]. Table 1 shows the accuracy measurement for data from the previous work.

Table 1. Types of cases for image comparison.

Category	Method's name	Technique used	Accuracy
Hybrid Approach Zhang et al. (2015)	LMS	• Lexicon based approach Unsupervised Learning	85.4%
Hybrid Approach Mudinas et al. (2012)	pSenti	• Lexicon based approach • Support Vector Machine	82.30%
Hybrid Approach (Khalifa & Omar, 2014)	Opinion QA	• Lexicon based approach • Naïve Bayes	91.0%
Hybrid Approach (Khalifa & Omar, 2014)	Opinion QA	• Lexicon based approach • Support Vector Machine	87.0%
Naïve Bayes (Khalifa & Omar, 2014)	Opinion QA	• Naïve Bayes	82.0%
Support Vector Machine (Khalifa & Omar, 2014)	Opinion QA	• Support Vector Machine	80.0%

From Table 1, LMS technique by using a machine learning classifier was trained to allocate polarities to sentiments and instead of the polarities being allocated manually, result of lexicon-based approach is used as a training data for the classifier [24]. In another work [23], a hybrid approach called pSenti is proposed, in which it uses the data collected from reviews. It calculates the weight of the sentiments, the rate of occurrence for adjectives and lastly, compares the result by using cross style setting technique. Khalifa & Omar [25] propose an Opinion Question Answering (Opinion QA) that uses a hybrid method to examine opinions on a particular product or service in Arabic language. A comparison of accuracy measurement between three different hybrid approaches are performed which include lexicon-based method with SVM, lexicon-based method with Naïve Bayes and lexicon-based method with K-Nearest Neighbor (KNN). Their results show that a combination of lexicon-based method with Naïve Bayes obtained the highest accuracy.

Data or information visualization is vital to users since it transforms the data into interactive visual illustrations [26]. Despite of the amount of information obtained, cutting edge analysis and high performance information visualization is one of most ideal approaches to discern vital relationships [27]. Different circumstances that show different measure of understanding use different visualization methods [28]. Text is often used for visualization as it has turned into a developing and progressively vital subfield of data visualization using various text mining algorithms [29]. Some of the visualization techniques that can be used for text visualization are scatter plot visualization, network visualization and word cloud visualization. Hence, this work performs the sentiment analysis towards some crime issues by visualizing the texts used by the public in the Facebook comments. The Malay texts clustering are visualized based on a combination of lexicon-based approach and Naïve Bayes, with the word cloud representation to visualize the results.

3. Methodology

This work applies the iterative modified waterfall model framework. In order to visualize the information obtained from Facebook based on crime news, several processes are involved which are data extraction, pre-processing of data, clustering and visualization. The overall process is described in this Section.

3.1. System Design Overview

After specifying all requirements and performed the preliminary study on related works, suitable methods are selected to be applied in data processing. Data are collected by extraction using the Facebook API. Next, filtering of texts is performed during data preprocessing where tokenization and stop word removals are involved. The chosen hybrid method of lexicon-based approach and Naïve Bayes are then utilized for text classification to analyze the sentiments. Word cloud is selected for visualization. The flow of the system is illustrated in Figure 1.

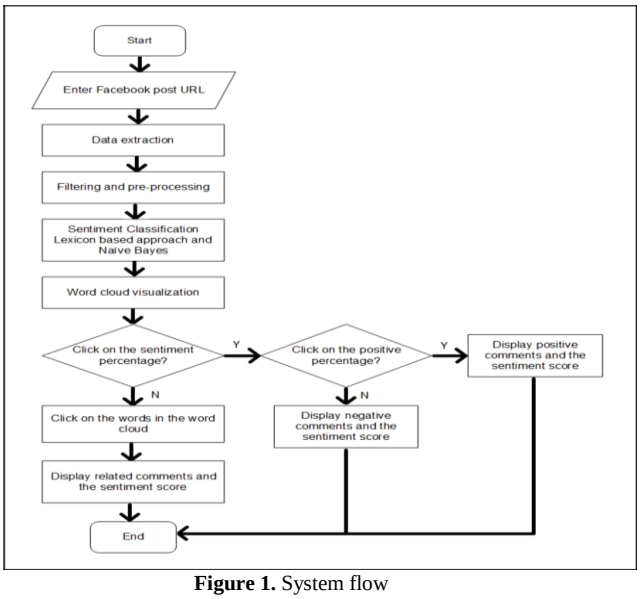


Figure 1. System flow

Based on Figure 1, data extraction is performed by extracting all comments from a crime news page. This page is obtained from the URL input by the user. Facebook API is used to retrieve the data in JSON format in order to clean the raw texts during data preprocessing phase. Then, filtering is done in order to filter the entire unnecessary images and symbols in the comments. This work only focuses on obtaining the texts without any symbols using tokenization and stop words removals. The cleaned text is used to calculate the number of occurrences for each meaningful word. The word occurrences determine word size to be visualized. The higher the occurrence, the bigger the font size of the word. The hybrid technique involving lexicon-based and Naïve Bayes are applied to train classifier for clustering the significant information.

3.2. Classification: Lexicon-based Approach and Naïve Bayes

In lexicon-based approach, a dictionary-based is used where it contains all the words that represent emotions. Next, the system performs the comparison between the words in the dictionary with the words in the comments to find a match. A trained classifier (Naïve Bayes) calculates the probability of word occurrences.

After training the datasets, the classification is performed by identifying the sentiments of each comment. The score is then calculated to determine the percentage of sentiment. In the following process, the comments are separated into many lists to allow the system to visualize various categories of comments which produce a more accurate result. Then, a word cloud presentation produces a suitable interactive visualization as it provides informative understanding.

3.3. Data Visualization: Word Cloud

Displaying most occurrence words in a textual document is called word cloud. The arrangement of words is random and in any format. The size of the word depends on its frequency in the document. Thus, the more frequent the words are used, the bigger size it becomes. Word cloud is the most suitable text visualization as compared to scatter plot and network visualization since it is able to show the words used in positive and negative sentiments. Words in scatter plot are typically represented by other symbol, so it would be troublesome for people to identify which words are being represented. On the contrary, the network visualization focuses more on the relationship between each word. Therefore, word cloud is the most appropriate as it focuses more on the usage of words. Figure 2 depicts the idea of how the output is presented for Malay language texts.

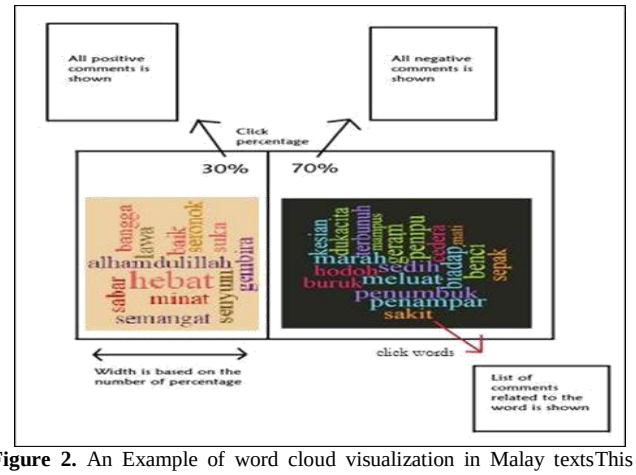


Figure 2. An Example of word cloud visualization in Malay textsThis representation assists users to analyze the sentiments of the selected issue. Each sentiment is represented by a set of words. The words can be clicked by the user to observe the visualization for each comment that is related to selected word. After undergo all the required processes, results are analyzed by observing the visualized output. The words that represent sentiments and other significant words are clustered into two categories which are positive and negative.

4. Results and Analysis

This section presents the user interface of the system by showing the output of each process. The first output displays the initial user interface for data extraction process. After filtering the extracted data, classification is performed to obtain the percentage of positive and negative sentiments. Lastly, the visualized output is presented for analysis.

4.1. User Interface

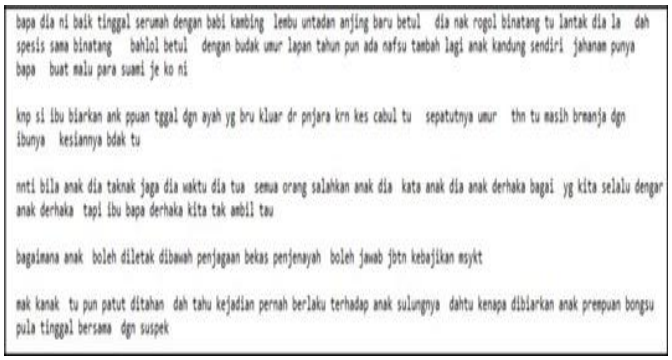
User interface is an important element in the system development. It consists of six main components: an input text field to retrieve the Facebook page URL, two text areas which display the filtered extracted comments (on the top) and words with their occurrences (at the bottom text area), and three buttons to run the system. The buttons are *Extract Data* which extracts all the comments from the Facebook page (page URL entered by the user), *Clear Data* to clear all the information on the interface, and *Submit* button to redirect the user to the panel showing the word cloud visualization. Figure 3 illustrates the user interface.

Figure 3. System user interface

Based on Figure 3, the data will be extracted when the user provides a Facebook page URL and the *Extract Data* button is clicked. The information displayed on the interface allows users to gain useful information about a crime issue.

4.2. Data Collection

Data for the system can be collected from a selected Facebook page that contains a post regarding crime news. Posts with most comments are selected as they help in creating more efficient word cloud. All comments in the selected post are extracted as raw data. These data are then filtered by removing all the unnecessary words and symbols, so that a user can focus on the important parts of the issue being discussed. The cleaned texts are then displayed onto the interface for the user as shown in Figure 4.


bapa dia ni baik tinggal serumah dengan babi kambing lembu untadan anjing baru betul dia nak rogol binatang tu lantak dia la dah spesis sama binatang bahlol betul dengan budak umur lapan tahun pun ada nafsu tambah lagi anak kandung sendiri jahanam punya bapa buat malu para suami je ko ni

knp si ibu biarkan ank ppuan tggal dgn ayah yg bru kluar dr pnjara krn kes cabul tu sepatutnya umur thn tu masih bnenja dgn ibunya kesiannya bdak tu

nnti bila anak dia taknak jaga dia waktu dia tua semua orang salahkan anak dia kata anak dia anak derhaka bagai yg kita selalu dengar anak derhaka tapi ibu bapa derhaka kita tak ambil tau

bagaimana anak boleh diletak dibawah penjagaan bekas penjenayah boleh jawab jbtn kebajikan msykt

nak kanak tu pun patut ditahan dah tahu kejadian pernah berlaku terhadap anak sulungnya dahtu kenapa diblarkan anak preepuan bongsu pula tinggal bersama dgn suspek

Figure 4. Facebook comments on the crime news that have been cleaned

Besides displaying the filtered data, the number of occurrences for each unique word available in the comments is displayed in order to observe the most frequent used words for the crime issue. This is important because the repeated word indicates the popular topic discussed by the public. It helps in solving a related matter. Next, the clustering and calculations of percentage for the sentiments are performed.

4.3. Classification Results

In this process, classification of sentiments is performed on each of the comment. The first classification is done by using the lexicon-based approach. Each word in the comments is checked by referring to the emotion library in the database. When a matched is found, the system updates the weight of each word according to the number of its occurrences. Next, the Naïve Bayes is performed to retrieve the words from the database and train them for classifi-

cation. The percentages of sentiments are then calculated. The word cloud displays the positive and negative words collected from the comments with the percentage values. The output is shown in Figure 5.

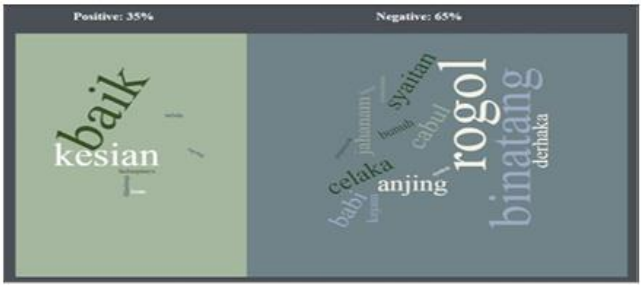


Figure 5. Word cloud visualization

From Figure 5, it shows the useful words being used in the comments related to a crime issue in a form of word cloud. In addition, the percentages of positive and negative sentiments are also displayed at the top of each category. The visualization panel is interactive as the user can click on the percentage to view the related comments either positive or negative.

Furthermore, the words displayed in the cloud can also be clicked by the user to view the comments posted by the public. Figure 6 illustrates the output when the user clicked on the words inside the word cloud.

Figure 6. Comments posted by public based on the words clicked

Hence, several experiments are conducted by observing two different case studies on crime issues which are Nhaveen bullying case, and Universiti Pertahanan Nasional Malaysia (UPNM) case. The case studies are visualized for analysis.

4.4. Visualization and Analysis

The first topic is the Nhaveen bullying case. The public sentiment can be observed from the sentiment percentage and the word cloud and the result is displayed in Figure 7.

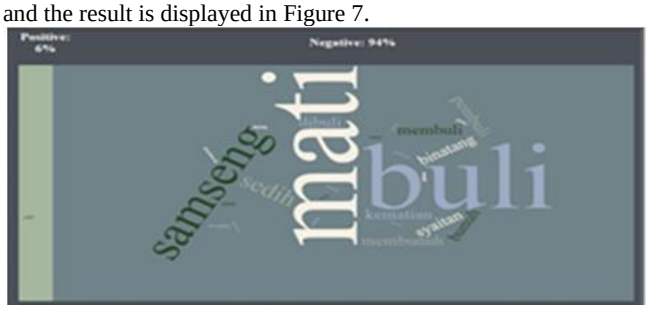


Figure 7. Nhaveen bullying case result

Based on Figure 7, the positive sentiment has a percentage of 6% whereas the negative sentiment has a percentage of 94% which shows that the public does not respond well to the crime news. The public is not satisfied with the crime news and cannot accept them well.

The second topic is the UPNM case on bullying and death. The visualized data is illustrated in Figure 8.

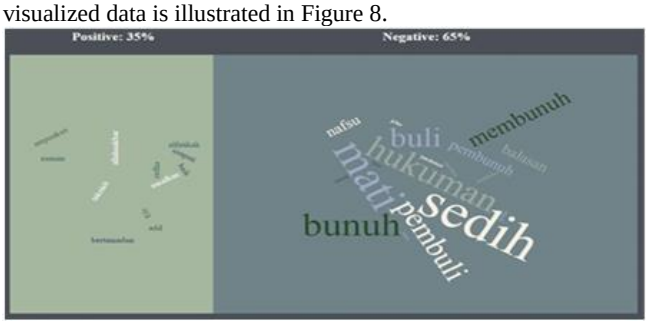


Figure 8. UPNM bullying and death case result

Based on the result, the percentage differences between the positive and negative sentiment is about 30%. The negative sentiment percentage is higher than the positive sentiment. It can be observed that most people are not satisfied with the crime news, but there are also a few who are satisfied and reacted well to the news.

5. Discussion and Conclusion

In this work, textual visualization is presented on a word cloud technique to present an informative public opinion. The use of different font sizes allows the results to be interpreted better. The larger the font size, the more of the word being used in the Facebook comments. Besides displaying the word group, the percentage of positive and negative sentiments are also important to determine people's reactions. The implementation is based on a hybrid method which is lexicon-based approach for word matching, and Naïve Bayes to train data on calculating word occurrences. Based on the two selected case studies, the results are more inclined to negative reactions. For a detailed observation on the chosen topics, the user is allowed to click on the percentage of positive and negative sentiments to view the comments where the texts have been cleaned. Moreover, the related comments can be viewed by the user when each of the word in the cloud is clicked. Such interactive feature allows more impact for users to do their analysis.

References

[1] M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53(1), 59–68 (2010).

[2] Mizuko, S. Baumer, M. Bittanti, D. Boyd, R. Cody, R. H. Stephenson, H. A. Horst, P. G. Lange, D. Mahendran, K. Z. Martínez, C. J. Pascoe, D. Perkel, L. Robinson, C. Sims and L. Tripp. Hanging out, messing around, and geeking out: Kids living and learning with new media. MIT press (2009).

[3] K. M. Lau, W. K. Hou, B. J. Hall, D. Canetti, S. M. Ng, A. L. F. Lam and S. E. Hobfoll. Social media and mental health in democracy movement in Hong Kong: A population-based study. Computers in Human Behavior, 64, 656–662 (2016).

[4] Paris, H. Christensen, P. Batterham and B. O'Dea. Exploring emotions in social media. Collaboration and Internet Computing, 54–61 (2015).

[5] T. J. Lampoltshammer, O. Kounadi, I. Sitko and B. Hawelka. Sensing the public's reaction to crime news using the "Links Correspondence Method". Applied Geography, 52, 57–66 (2014).

[6] M. Mohd and N. M. Ali. An interactive Malaysia crime news retrieval system. International Conference on Semantic Technology and Information Retrieval, STAIR 2011, 220–223 (2011).
Perikos and I. Hatzilygeroudis. Recognizing emotions in text using ensemble of classifiers. Engineering Applications of Artificial Intelligence, 51, 191–201 (2016).

[7] N. Tsirakis, V. Poulopoulos, P. Tsantilas and I. Varlamis. Large scale opinion mining for social, news and blog data. Journal of Systems and Software, 127, 237 – 248 (2016).

Przepiorka, A. Blachnio and J. F. Diaz-Morales. Problematic Facebook use and procrastination. Computers in Human Behavior, 65, 59–64 (2016).
P. Pillai, A. Ashok and S. Gunasekar. Facebook as a site of information: A study in the Indian context. IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2015, 1-4 (2016).
[8] Alarifi, M. Alsaleh and A. M. Al-Salman. Twitter turing test: Identifying social machines. Information Sciences, 372, 332–346 (2016).
[9] P. Sheldon and K. Bryant. Instagram: Motives for its use and relationship to narcissism and contextual age. Computers in Human Behavior, 58, 89–97 (2016).
[10] K. Iliakopoulou, S. Papadopoulos, and Y. Kompatsiaris. News-oriented multimedia search over multiple social networks. International Workshop on Content-Based Multimedia Indexing, 1-6 (2015).
[11] S. Veeramani, and S. Karuppusamy. A survey on sentiment analysis technique in web opinion mining. International Journal of Science and Research (IJSR), 3(8), 1776–1780 (2012).
[12] S. Roy, S. Nag, I. K. Maitra and S. K. Bandyopadhyay. A Review on Automated Brain Tumor Detection and Segmentation from MRI of Brain. International Journal of Advanced Research in Computer Science and Software Engineering, 3(6), 594–598 (2013).
[13] G. Vashisht and S. Thakur. Facebook as a Corpus for Emoticons-Based Sentiment Analysis. International Journal of Emerging Technology and Advanced Engineering, 4(5), 904–908 (2014).
[14] E. R. Tufte. The Visual Display of Quantitative Information. Technometrics, 197 (2001).
Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. Exploiting emoticons in sentiment analysis. Proceedings of SAC'13, the 28th Annual ACM Symposium on Applied Computing, 703–710 (2013).
[15] S. Park, and Y. Kim. Building thesaurus lexicon using dictionary-based approach for sentiment classification. IEEE/ACIS 14th International Conference on Software Engineering Research, Management and Applications, SERA 2016, 39–44 (2016).
[16] S. Singh, and T. J. Siddiqui. Evaluating effect of context window size, stemming and stop word removal on Hindi word sense disambiguation. International Conference on Information Retrieval and Knowledge Management, CAMP'12, 1–5 (2012).
Ekawijana and H. Heryono. Composite Naive Bayes Clasification and semantic method to enhance sentiment accuracy score. 4th International Conference on Cyber and IT Service Management, 1–4 (2016).
[17] G. Vaitheeswaran. Combining Lexicon and Machine Learning Method to Enhance the Accuracy of Sentiment Analysis on Big Data, 7(1), 306–311 (2016).
Mudinas, D. Zhang and M. Levene. Combining lexicon and learning based approaches for concept-level sentiment analysis. Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12, 1–8 (2012).
[18] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu and B. Liu. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89, 1–8 (2015).
[19] K. Khalifa and N. Omar. A hybrid method using lexicon-based approach and Naive Bayes classifier for Arabic opinion question answering. Journal of Computer Science, 10(10), 1961–1968 (2014).
[20] S. Liu, W. Cui, Y. Wu, M. Liu. A survey on information visualization: recent advances and challenges. Visual Computer, 30(12), 1373–1393 (2014).
[21] SAS.: Data Visualization Techniques. Retrieved from http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-visualization-techniques-106006.pdf. (2013).
[22] M. Khan, and S. S. Khan. Data and Information Visualization Methods, and Interactive Mechanisms: A Survey. International Journal of Computer Applications, 34(1), 975–8887 (2011).
[23] K. Kucher. Text Visualization Browser : A Visual Survey of Text Visualization Techniques. InfoVis (2014).