Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 59 (2015) 373 – 381

Procedia Computer Science

International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)

Collaborative Social Network Analysis and Content-based Approach to Improve The Marketing Strategy of SMEs in Indonesia

Warih Maharani, Alfian Akbar Gozali*

School of Computing, Telkom University Bandung 40257, Indonesia

Abstract

Social Network Analysis (SNA) has been applied in several case studies. SNA is applied to enhance the company's marketing strategy as well as small and medium businesses. This research proposes a collaborative model using content-based and user-based approach, with the centrality measurement methods. Content-based approach tends to focus on tweet content analysis of the existing nodes in a network, while the user-based approach focuses on the connections between users in the network twitter. The model will combine the advantages of collaborative content-based and user-based approach, to find the most influential people in a twitter network to make the dissemination of information more effectively and efficiently.

Peer-review under responsibility of organizing committee of the International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)

Keywords: Social Network Analysis (SNA); centrality measurement; content-based; user-based

1. Introduction

Along with the development of social networks, research on Social Network Analysis (SNA) are now being applied in a variety of case studies, such as for choosing the most influential politician in a country, for human identification affected by HIV in a population, to build an interactive map, or to choose the most popular employees in the company. Social networking is built on the idea that there is a determinable structure to how people know each

* Corresponding author. Tel.: +62-22-7564108
E-mail address: wmaharani@telkomuniversity.ac.id

Peer-review under responsibility of organizing committee of the International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)
doi:10.1016/j.procs.2015.07.540

other, whether directly or indirectly. Wasserman (11) states that SNA is based on an assumption of the importance of relationships among interacting units. While Freeman (12) defines the SNA as a technique that focuses on the study of human interaction patterns that do not appear explicitly. Scott (1) defines a set of methods to investigate aspects of the relations in the social structure, methods which are specifically geared towards an investigation of the relational aspects of these structures. The use of these methods, therefore, depends on the availability of relational rather than attribute data. Based on these definitions, broadly have the same meaning, which leads to a process of social network analysis deals with the structure and patterns of interaction entities in it. So, SNA is the mapping and measuring of relationships and flows between people, groups, organizations, computers, web sites, and other information/knowledge processing entities. The pattern of interactions between entities will provide new information. Entities will become nodes in a graph that has information to help the researcher in making hypotheses on the phenomena that occurs.

SNA describes the social relations of nodes in a network and ties are often called the edge, links, connection (5,12). In SNA, the value of influential nodes in a network is called centrality (12). Centrality is a value that describes how many connections from one node to other nodes (5,12). There are various ways that are often used to define centrality in a network to identify the influence of each node : degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality (12,13). The most influential nodes are the nodes that have the greatest weight value, both in terms of the number of interactions or the number of connections between users. The basis of the determination of the most influential nodes in a network is the centrality measures.

Nevertheless, SNA tends to focus more in terms of structure and patterns of interaction between entities. Freeman (12) stated that the pattern of human interaction is an important aspect of human life involved. Interaction in SNA will be able to answer a variety of issues, among others, to measure how individuals connect with others, how one person will affect the relationships between people and measure how individuals in a group are connected and interact. Previous studies have focused on the number of nodes with the weighting between nodes, however do not consider the content of the information conveyed in the network, but only focus on the interaction relationship of a node with other nodes.

The objective of this paper is to propose an architecture of collaborative approach, by applying some centrality measurement approaches to find the most influential users in Small and Medium Enterprises (SMEs) network. The proposed model will consider the content and relationship between nodes in a network with a certain weighting methods. How to analyze the structure of the social media to find the most influential users on Twitter, so it can be used to accelerate the dissemination of information in the marketing of products or services of SMEs in social media.

The paper is organized as follows : Section 2 presents the related works of this paper ; Section 3 describes the research methods ; Section 4 discusses the results and discussion ; finally, Section 5 concludes the paper and challenges as well as opportunities for further research.

2. Related Works

In general, the main goal of SNA is to identify and calculate the “central” nodes in a network. The “central” nodes is important because it represents the critical position in the network, which correlated with nodes popularity. Centrality analysis is widely used in the SNA to analyze the ability of each node in the flow of information to determine centering graph. There are various applications that implemented SNA, such as SNA is used to calculate how well used a road is in a transportation network, how important a web page or how important a room is in a building. The most widely used centrality measurement are degree, closeness, betweeness and eigenvector centrality (12,13).

User-based approach is more focused on the similarity between users based on the number of relationships. This approach does not consider the content of nodes in a network. This approach can not represent the content of interaction that occurs between nodes because it only considers the connectivity between nodes. For example, user-based approach in twitter network only consider number of relationship of central nodes, rather than content in interaction between nodes such as retweet, reply, follow or mention interactions.

Previous studies have examined correlation among centrality measures (31,32). Kretschmer proposed a new centrality measure for social network analysis applicable to bibliometric and webometric data (2). Maharani (31) compared degree centrality and eigenvector centrality to find the most influential users in twitter. The result showed

that both centrality can produce different nodes that represent the most influential users in twitter network. In another study, Zudha (32) implemented Kretschmer methods to find the most influential users in a network. Another studies performed community detection with clustering approach with agglomerative/hierarchical, partition-based and laplacian graph (3,4). However, these approaches have drawbacks, which only consider the structure and user interaction in a network. In addition, they have not considered the information content and its distribution on the web as a whole and have not used metadata and semantic features.

2.1 Degree Centrality

Degree centrality is defined as the number of degrees of relationship that is connected to a node, for example in a relationship of friendship in a community (12). The central actor is one with many connections. In this research, people who are considered the most popular person is one who has the most friends relationship, so that the value of the measurement for the degree centrality is calculated based on the number of friendship connections.

If there is a graph with n nodes, then the degree centrality is defined in the formula (5,12) below :

$$C_D(V_i) = \sum_{k=1}^{n} a(V_i, V_k) \quad (1)$$

Where a(u,v) = 0 if u and v not connected by a relationship, otherwise a(u,v) = 1.

2.2 Closeness Centrality

Closeness centrality is a measure that is based on geodesic distances, which measures how many node to another node. The central node is one that is close to other nodes in a network. Measurement of the closeness centrality indicates availability, safety and security (12). Closeness centrality can be measured by the formula :

$$C_C(V_i)^{-1} = \sum_{k=1}^{n} d(V_i, V_k) \quad (2)$$

Pan (6) stated that according to many researchers, closeness centrality is not always appropriate in some cases, especially if the network is to be measured are large. It is due in large networks, the tendency of the proximity of a node only on the small scope of the overall large graph, causing the value of closeness for each node is small. The small value of closeness does not provide much information, because the actual values of the geodesic distances for each node can represent more information. Value-added operations to seek closeness, will eliminate a lot of information. In the case of a large network, closeness value must be accompanied by the value of the geodesic distances vector. So, both parameters, closeness and geodesic distances vector, will be representing the value of centrality in the network.

Different methods and algorithms have introduced to measure closeness, like the random-walk centrality. Noh (7) introduced that there is a measure of the speed which randomly walking messages reach a vertex from elsewhere in the network, a sort of random-walk version of closeness centrality. Stephenson and Zelen (8) proposed closeness measure, which bears measures the harmonic mean length of paths ending at a vertex, which is smaller if it has many short paths connecting it to other vertices. Dangalchev (9) modified the definition for closeness so it can be used for disconnected graphs and the total closeness is easier to calculate. Another extension to networks with disconnected components has been proposed by Opsahl (10).

2.3 Betweeness Centrality

Betweeness centrality is a measure that focuses on the ability of a node in a position to connect the nodes to one another via the shortest path through it. So, this approach measures the number of shortest path that passes through the node. The more times a node lies on the shortest path between two other nodes, the more control that the node has over interaction between these two non-adjacent nodes (11). A node with the highest betweeness centrality can

influence the other nodes in a network, by communicating between others (12). Betweeness centrality formulated as follows (12) :

$$C_B(V_i) \quad = \sum_{k=1}^{n} \sum_{k=1}^{j-1} g_{jk}(V_j)/g_{jk} \tag{3}$$

Where g_{jk} is the total number of geodesic path (shortest path) that connects V_i and V_k , and g_{jk} is the number of geodesic path that passes through V_i. The algorithm for calculating the betweeness centrality is very complex and requires a lot of memory allocation, which can achieve a complexity of $O(n^3)$ (12).

2.4 Eigenvector Centrality

Eigenvector centrality is a measure of a most influential nodes in a network. Eigenvector centrality value relative to all nodes in the network based on the principle that the node with the highest score is the node that neighbors have Eigenvector centrality great value (13). A central actor is connected to other central actors.

$$C_i^e = \frac{1}{\lambda} \sum_{j:j \neq i} y_{i,j} c_j^e \tag{4}$$

2.5 Content-based Approach

Content-based approach with regard to opinion mining research, which involves the opinion extraction and classification of social media content. Content on social media can include several components, among other objects, features objects, opinion terms, opinion holder and time (14). The fifth component is generally related to the SNA, where the SNA will seek relationship linkages among the five components are connected by nodes that exist in the SNA. There are two approaches that are generally used in the extraction process, which is based on natural language processing and machine learning approaches. Several studies using natural language processing techniques to identify the object, the object features and opinion (14,20). In general, some studies using the Part-of-Speech (POS) tagging and parsing the syntax tree to identify the noun and noun phrase (15–18). Zhuang (20) using WordNet as well as knowledge of the movie dataset for extracting movie features and its opinions. In general, POS tagging and parsing technique has the advantage that a relatively simple calculation, as well as the average yield relatively high accuracy (17,18). However, this technique can not find all the objects in the document, because not all objects represented in noun phrases.

The other task in content-based approach is polarity classification. Polarity classification determines the opinion orientation of an object into positive and negative opinion (18–20). There are two approaches that are generally performed in classifying opinions, such as machine learning approach and lexicon-based approach. Several studies used supervised learning techniques (26,28), a semi-supervised learning (21,22), and unsupervised learning (23,24). Supervised learning method that is widely used in opinion is Naive Bayes classification (25,27,28), Maximum Entropy (26,27), Artificial Neural Network (ANN), and Support Vector Machine (SVM) (28,29). Content-based approaches focus on content similarity between users, both in terms of content as well as the proximity of keywords polarity opinion content. Inter-node is said to have a high similarity if it has a high content closeness.

2.6 Collaborative Approach

Content-based approach has advantages by considering the content in a network, while having a weakness because it does not consider the number of relationship connectivity between nodes directly. Collaboration between content-based approach and user based can combine the advantages of both approaches and cover their weaknesses. Therefore, by utilizing the similarity relationships among nodes and content similarity occurs between nodes, can further represent the real conditions that occur in social media and can be used to enhance marketing strategy in SMEs.

2.7 Small and Medium Enterprises

According to Indonesia's Law No. 9 In 1995, small businesses are small-scale productive activities and meet the criteria of net wealth at most two hundred million rupiah, excluding land and buildings or having sales at most one billion rupiah per year and can receive credit from the bank above the maximum fifty million rupiah to five hundred million rupiah. Indonesia is a developing country, so it needs to be advanced further industrial SMEs that represent economic progress in Indonesia. Moreover, with the global market in Indonesia, will increasingly force SMEs to further develop the SMEs industry.

One of the main efforts to develop SMEs is to provide an alternative way to disseminate information about the products of SMEs which currently tend to compete when compared with foreign products. For that, with the use of technology and the rapid development of social media, can be used to assist in disseminating information on SMEs products more easily, effectively and efficiently.

3. Research Method

3.1 Motivation

Social media has been widely used in Indonesia as a way to socialize, for the purposes of political campaigns, marketing, e-commerce, news distribution, social control as well as a medium of interaction with fans for the public figure. It certainly can give a positive or negative impact on the social environment.

Indonesia as a developing country have various SMEs that need to be developed. The positive effect in the presence of social media can be used to help develop SMEs in Indonesia, particularly in terms of the spread of information through social media. With the widespread dissemination of information and quickly through social media is expected to help SMEs. How to determine the structure of social relations that exist in the media so that it can be seen that the most influential users in Social Media that can be used to accelerate the dissemination of information in marketing products and services of SMEs in social media.

3.2 Challenges

Some challenges in modeling centrality measurement in SNA with the collaboration of the content and user-based approaches are:

Challenge 1 :

How to determine and collaborate centrality measurement and content-based approach using degree centrality, closeness centrality, betweeness centrality and eigenvector centrality in twitter for SMEs. Each approach can be used to analyze network on twitter with some advantages and disadvantages, but it need a comprehensive analysis to determine and collaborate the approaches to determine the most influential people on twitter network.

Challenge 2 :

How to determine the tuning parameter on the centrality measurement with content-based and user-based to accommodate problems in SNA tweet content, as well as any interaction problems between connected nodes?

Current SNA research focuses on the centrality measurement to get the most influential nodes on a network. While current opinion mining research focuses on opinion extraction and classification. Both of these studies have relevance in terms of the interactions that occur in a social network, so that the collaboration between these approaches should be done so that it can generate a representative analysis with a combination of interaction and tweet content. The collaboration can implement using tuning parameter to obtain the best performance between content-based and user based approaches.

Challenge 3 :

How to determine the most influential user, based on collaborative model using content based and user-based approach in SME as the case studies?

The previous research showed that centrality weighting method affect the results of the most influential user on twitter (31,32). However, such approaches do not consider the semantics of tweets that occur in the network. Hypothesis with the semantic analysis of content-based approach, will be representing the connections that occur in the network, which can lead to better system performance.

Challenge 4 :

How most influential users can improve the marketing strategy for SMEs?

SNA has been applied in several fields, among others, as one of the main marketing strategies. However, the majority of businesses in Indonesia engaged in small and medium enterprises. How to improve the marketing strategy for SMEs will be the main objective of this research, and then analyzed the impact of the most influential user to increase SME marketing.

Challenge 5 :

How to measure performance based on the most influential users for improving SMEs marketing strategies? Another problem is how to evaluate the performance of collaborative SNA and performance of SMEs marketing strategy. System performance can be measured with quantitative and qualitative approaches. While the performance of the marketing strategy requires evaluation approach involving business strategy that involves another disciplines.

4. Results and Discussion

Based on the above objectives and opportunities, this research propose a SNA collaborative model with content-based approach and user-based approach for SMEs in Indonesia.

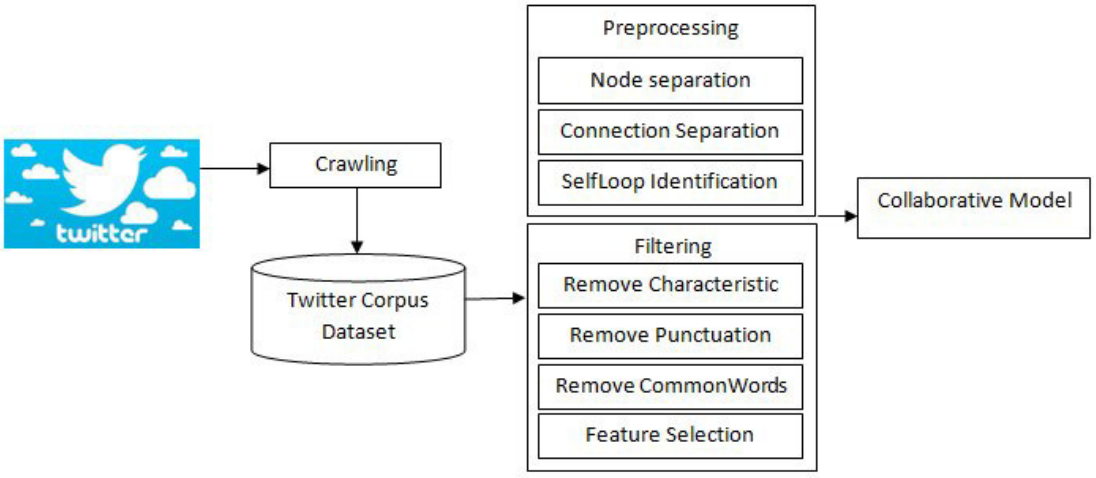

Fig. 1. General System Design

Figure 1 shows the proposed the collaborative model of content-based and user-based approach. The system consists of several tasks : (1) crawler, (2) preprocessing, and (3) collaborative model. Tweet Data retrieved through the crawling process, which represents the user and keywords of an SMEs. Crawling process includes crawling nodes, tweet content and other attributes required include crawling time and location. While in preprocessing, conducted node separation, connection separation, self-loop identification and filtering. In final step (3), the result will combine content and user similarity using weighting collaboration methods.

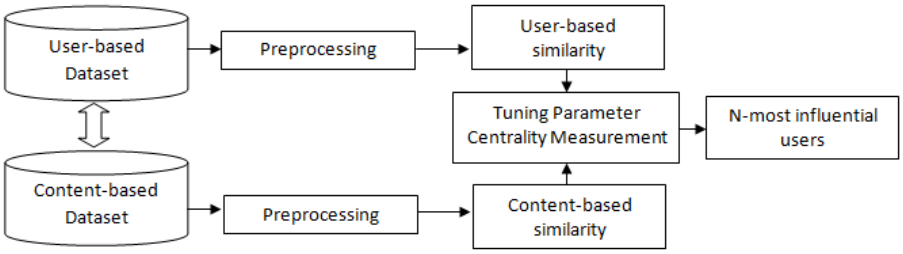

Fig. 2. Collaborative Model

The crawling process will be carried out to obtain the data relationship that are established between the twitter user whether it was followed, follower, mention, or reply relationship. In addition to the data relationship, also performed the data content crawling tweets that occur in the network that will be analyzed based on certain keywords. Preprocessing is the early stages of processing which aims to manage the dataset and analysis before it is processed by the system. In general, the preprocessing consists of two parts, namely parsing which is the separation between nodes and their relationships, and creating a table that contains the nodes and weight relationships. The text preprocessing is needed after the filtering such as stopword removal and stemming.

The goal of preprocessing is to create a (n x n) matrix or adjacency matrix to build a graph. While the second preprocessing is to choose a feature of existing tweet contents. The matrix contains the weight of the relationship that occurs between nodes. The content similarity can be calculated using syntactic similarity or semantic similarity. This paper proposed a collaborative approach which combine content-based and user-based approach. The collaborative approach requires a tuning parameters, which can adjust the weight of each approach as shown in Figure 2. The influence of their content and relationships can be set by using this weights. Relations that occur between nodes not only consider the degree of relationship, but covers the interactions that occur between nodes. For the case of Twitter dataset, the interaction that occurs in the form of following action, followed, mentions, reply and retweet. Each tweet is considered as micro document distributed to various users, so any additions tweet flowing in the friendship relations between users, whether the result of the retweet, reply or mention could be considered as weights. For each of these interactions always involve one tweet as the smallest unit, in a sense, for every mention, reply and retweet can not be done in 2 pieces tweet or more at once. So if the weight measurement process based on the number of tweets, the interaction mentions, reply and retweet it will have a weight of 1 unit.

Each of these interactions can affect the system performance, as seen in (30–32). While the content-based approach, can be done with the keyword-based, syntactic-based and semantic-based (30). Keyword-based approach will only consider the proximity of keywords that occur between nodes, while the syntactic-based approaches will consider the emerging patterns between nodes. Semantic-based approach will better utilize natural language processing to be more representative in understanding the semantic meaning that occurs between nodes.

Collaboration between keyword-based approach, syntactic and semantic, with the relationship-based approach, will produce a structure of social network analysis more representative. The model will be used as a marketing strategy for SMEs to market their products more effectively and efficiently. The proposed method has been supported by the results of the initial experiments that we have done in (31,32). The results showed that the centrality measure and the interaction that occurs between nodes, affect the determination of the most influential nodes in a network twitter.

5. Conclusion and Future Works

With the rapid growth development of social media, can be utilized for SMEs in Indonesia to improve their marketing strategies. Through the N-most influential user, SMEs can disseminate information products and marketing more effectively and efficiently. This research proposes a collaborative method by combining content-based and user-based approach to produce the most influential twitter user in the network. The proposed method involves tuning parameters that will affect system performance.

Future research will implement this proposed method using SMEs tweet dataset that will produce N-most influential user in their network. So, using those result, SMEs can disseminate their product information quickly and effectively.

References

1. Scott J. Social network analysis. Sage; 2012.
2. Kretschmer H, Kretschmer T. A new centrality measure for social network analysis applicable to bibliometric and webometric data. Collnet Journal of Scientometrics and Information Management. Taylor & Francis; 2007;1:1–7.
3. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L. Hierarchical organization of modularity in metabolic networks. science. American Association for the Advancement of Science; 2002;297:1551–1555.
4. Flake GW, Lawrence S, Giles CL. Efficient identification of web communities. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2000. p. 150–160.
5. Gaskins R. Network dynamics in saga and society. Scandinavian Studies. JSTOR; 2005;201–216.
6. Pan L. Effective and efficient methodologies for social network analysis. Virginia Tech; 2007;
7. Noh JD, Rieger H. Random walks on complex networks. Physical review letters. APS; 2004;92:118701.
8. Stephenson K, Zelen M. Rethinking centrality: Methods and examples. Social Networks. Elsevier; 1989;11:1–37.
9. Dangalchev C. Residual closeness in networks. Physica A: Statistical Mechanics and its Applications. Elsevier; 2006;365:556–564.
10. Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks. Elsevier; 2010;32:245–251.
11. Wasserman S. Social network analysis: Methods and applications. Cambridge university press; 1994.
12. Freeman LC. Centrality in social networks conceptual clarification. Social networks. Elsevier; 1979;1:215–239.
13. Bonacich P. Some unique properties of eigenvector centrality. Social Networks. Elsevier; 2007;29:555–564.
14. Liu B. Sentiment analysis and opinion mining: synthesis lectures on human language technologies. Morgan & Claypool Publishers. 2012;
15. Popescu A-M, Nguyen B, Etzioni O. OPINE: Extracting product features and opinions from reviews. Proceedings of HLT/EMNLP on interactive demonstrations. Association for Computational Linguistics; 2005. p. 32–33.
16. Popescu A-M, Etzioni O. Extracting product features and opinions from reviews. Natural language processing and text mining. Springer; 2007. p. 9–28.
17. Hu M, Liu B. Opinion extraction and summarization on the web. AAAI. 2006. p. 1621–1624.
18. Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2004. p. 168–177.
19. Ku L-W, Liang Y-T, Chen H-H. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006. p. 100–107.
20. Zhuang L, Jing F, Zhu X-Y. Movie review mining and summarization. Proceedings of the 15th ACM international conference on Information and knowledge management. ACM; 2006. p. 43–50.
21. Zhai Z, Liu B, Xu H, Jia P. Clustering product features for opinion mining. Proceedings of the fourth ACM international conference on Web search and data mining. ACM; 2011. p. 347–354.
22. Goldberg AB, Zhu X. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing. Association for Computational Linguistics; 2006. p. 45–52.
23. Ganesan K, Zhai C, Han J. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics; 2010. p. 340–348.
24. Turney PD. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics; 2002. p. 417–424.

25. Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences. Elsevier; 2011;181:1138–1152.
26. Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and trends in information retrieval. Now Publishers Inc.; 2008;2:1–135.
27. Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics; 2002. p. 79–86.
28. Zhang L, Liu B. Identifying noun product features that imply opinions. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. Association for Computational Linguistics; 2011. p. 575–580.
29. Tan S, Zhang J. An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications. Elsevier; 2008;34:2622–2629.
30. Maharani W. Microblogging sentiment analysis with lexical based and machine learning approaches. Information and Communication Technology (ICoICT), 2013 International Conference of. IEEE; 2013. p. 439–443.
31. Maharani W, Gozali AA, others. Degree centrality and eigenvector centrality in twitter. Telecommunication Systems Services and Applications (TSSA), 2014 8th International Conference on. IEEE; 2014. p. 1–5.
32. Rachman ZA, Maharani W, others. The analysis and implementation of degree centrality in weighted graph in Social Network Analysis. Information and Communication Technology (ICoICT), 2013 International Conference of. IEEE; 2013. p. 72–76.