■ 1416

Selecting User Influence on Twitter Data Using Skyline Query under MapReduce Framework

Ahmad Luky Ramdani*[1], Taufik Djatna[2], Heru Sukoco[3]
[1]Department of Science, Informatic Engineering, Institut Teknologi Sumatera, Indonesia
[2]Department of Agro-industrial Technology, Bogor Agricultural University, Indonesia
[3]Department of Computer Science, Bogor Agricultural University, Indonesia
*Corresponding author, e-mail: taufikdjatna@ipb.ac.id[2], ahmadluky@if.itera.ac.id[1]

Abstract
The aim of this research was to select and identify user influence on Twitter data. In identification stage, the method proposed in this study was matrix Twitter approach, sentiment analysis, and characterization of the opinion leader. The importan characteristics included external communication, accessibility, and innovation. Based on these characteristics and information from Twitter data through matrix Twitter and sentiment analysis, a algorithm of skyline query was constructed for the selection stage. Algorithm of skyline query selected user influence by comparing with other users according to values of each characteristic. Thus, user influence was indicated as user that was not influenced by other users in any combination of skyline objects. The use of MapReduce framework model in identification and selection stage, support whole operation where Twitter had big size data and rapid changes. The results in identification and selection of user influence exhibited that MapReduce framework minimized the execution time, whereas in parallel skyline query could reveal user influence on the data.

Keywords: opinion leader, user influence, skyline query, mapreduce framework

1. Introduction

Twitter has been a popular media in Indonesia that is widely used for sharing opinion and propaganda. Based on motivation "what's happening?", The users deliver their opinion and propaganda with less than 140 characters as called by tweet. The propaganda-related tweet is commonly found in voters of a public election [1] and a product marketing [2]. In the public election, propaganda is addressed to national figures or prospective leaders, while, in marketing, it is related to product branding. The use of social media Twitter to blow up propaganda is called by viral marketing [3].

One of the concepts used in the viral marketing is user influence [4]. It is a user in Twitter that is capable to influence other users in making decision. Matrix Twitter (MT) is one of the methods used for impact measurement. MT is a mathematic symbol that provides information related to Twitter user in numerical values such as number of follower, following, reply, and tweet [5]. These values are combined in formula or algorithm like in the use of algorithm PageRank in selecting user influence [6]

The use of MT in identifying user influence was conducted by [7] based on characteristics of opinion leader and machine learning algorithm. However the study had a weakness; analysis of user influence was only based on MT. MT only measures users based on their information and relation to other users. Septiandri and Purwarianti (2013) stated that the influence level of users may result from their tweet. The reply sentiment in a tweet demonstrated the influence of the user [8].

Furthermore, machine learning algorithm highly depends on data used in the training and measurement, thus model consistency is always required in different various data. Different data possibly cause inaccuracy in identifying the user influence. [9] used skyline query to identify and select users in Facebook. Skyline query is an algorithm to obtain an object that is not dominated by other objects [10]. The object is called as skyline. The user influence is considered when the user is not dominated by other users. These users indicate better particular characteristics. Hence this research used MT feature and tweet content analysis to select user influence using skyline query based on characteristics of opinion leader (OL).

OL represents the individuals that are able to influence others in decision making of a community [11]. User influence in twitter has similarity to OL [12]. [13] stated that OL could be seen from characteristics including external communication, innovation, accessibility, and social-economy status.

Different from common skyline query, selection of user influence using OL characteristics is not simple. The characteristics are not directly from Twitter data. Increasing Twitter data, data processing is unable to be processed using simple approach in a computer. MapReduce in a group of computer is a solution for determination of user influence. It is a programming model and software that relates to big size data [14]. MapReduce framework can process distributed data in the group of computer. Based on the description, the study aimed to (1) identify OL based user influence according to MT and tweet content analysis in MapReduce framework, and (2) apply algorithm of skyline query in MapReduce framework to select user influence.

2. Research Method

The method consists of 3 stages, 1) pre-processing 2) identification of user influence, and 3) determination of user influence. Data collection based on topics and sentiment analysis are conducted in pre-processing, while selection of user influence includes skyline process and top-k query. Identification and selection of user influence are processed in MapReduce framework. The research flowchart is presented in Figure 1.
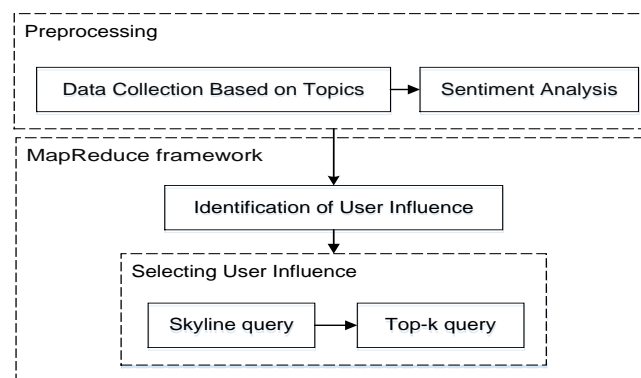

Figure 1. Research stages

2.1. Pre-processing
2.1.1. Data collection based on topics

This step identifies a character to obtain national figure for topic and keyword in data collection. This uses contextual and morphology on data article from online news media and channels of the political party such as detik.com. Article data was obtained by using Really Simple Syndication (RSS). Contextual and morphology rule are chosen since they have better accuracy compared to association mining rule in identifying the characters [15]. The characters obtained are then used to determine keywords. The data was collected using library Twitter4j. In this study, only opinion tweet is selected for next stage. Therefore, normalization and selection of opinion are performed. Part of speech tagging, Hidden Markov Model algorithm (HMM) and opinion sentence rule are used to filter the opinion [16]. This step results in Twitter containing opinion that comment national figures.

2.1.2. Sentiment analysis

The step aims to obtain model of sentiment classification. The model is used to determine the level of reply sentiment in user influence. The step employs hybrid, a combination of machine learning algorithm and sentiment dictionary (lexicons) to gain classification model [17].

The hybrid method consists of several process including corpus formation, normalization, feature reduction, standardization, mark extraction, and classification. Formation of corpus is based on [18-19], which uses emoticon characters. Normalization is to improve invalid data in Twitter [20]. Feature reduction aims to reduce unused word/characters such as punctuation and stop words. Standardization of sentiment dictionary consists of translation, validation, and standardization. Translation is used by means of library TextBlob in Bing Liu [21], SentiStrength Emoticons [22], SentiWords [23], and MPQA Subjectivity (MPQA-S) [24]. Validation employs Kamus Besar Bahasa Indonesia (KBBI) to make sure that translated output has meaning in Indonesia language. Furthermore, each dictionary has different sensitivity range. Therefore, standardization is applied in value 0 and 1 [25].

Standardized dictionary is used in mark extraction. This step is to determine vector of mark sentiment used in the sentiment classification. The marks used are from [26] with modification, namely, word number of positive, negative, and neutral sentiments, word number of sentiments, word value of sentiment, maximum value of positive and negative sentiments, and number of negation word in the tweet. The vector and machine learning algorithm, Naive Bayes Classification (NBC), are used in the classification to achieve classification model. The algorithm NBC provides high accuracy in the classification of text document [27]

2.2. Identification of user influence

Identification of user influence is based on specification and characteristics of OL, MT, and classification model of sentiment. Therefore, data attribute, MT calculation, and sentiment weight of user are analyzed. The proses uses BPMN workflow for analyzing and modelling a process. Programming language java and Hadoop are used for implementation. BPMN used is Sybase Power Designer 16.0. BPMN is understandable notation for analyzing, modelling, and representing process stream. The OL characteristics are external communication, accessibility, and innovation adaptation. Additionally, social-economy status is discarded since Twitter users have good social-economy status [7].

a. Analysis of data attribute

Attribute analysis aims to determine data attribute for calculation of MT value, which is conducted using data observation.

b. Calculation of MT value

Calculation of MT value is based on MT [5]. This step is needed since some values of MT are not directly obtained from Twitter data. The output of this process is Twitter user and MT value.

c. Calculation of sentiment weight

Calculation of sentiment weight uses model from sentiment analysis. The use of sentiment model only focuses on sentiment value of tweet, which is a reply. This is because value of reply sentiment on tweet shows user'ss influence [8]

2.3. Selecting user influence
2.3.1 Skyline query

Skyline query is a method to select objects that are not dominated by other objects in data. These objects are skyline. Skyline algorithm includes Block Nested Loops (BNL) [28], Sort Filter Skyline (SFS) [29], Nearest Neighbor Skyline (NN) [30], Bitmap, Index and Branch-and-Bound Skyline [31]. Sort Filter Skyline (SFS) algorithm is development of Block Nested Loops (BNL) algorithm.

Based on the value from identification, a user is skyline if and only if this user is not dominated by other users. User i is not dominated by other users when value of user I in each OL characteristics is better than other users. Therefore, a user influence is not dominated by other users based on OL characteristics. Skyline query in the current study uses algorithm Sort Filter Skyline (SFS) in MapReduce framework [32].

SFS in MapReduce framework consists of tow steps, in which determination of local and global skyline. Determination of local skyline is initially conducted by splitting data. Each part is counted to attain skyline based on SFS algorithm. Skyline obtained from each part is merged and filtered to gain global skyline. Partition process employs technique used in Nearest Neighbor and Divide & Conqure. Data is divided by 2^d parts based on median value from each data attribute, which d is number of data attribute. Each part is defined as d-bit, which is used in key-value of function map and reduce. The result of local skyline determination is user/skyline

influence from each part. The data were used in the determination of global skyline to obtain user influence from whole data.

2.4.2 Top-k query

Top-k query is ranking method for user influence, which k is a positive integer. The value defines expected number of the most influential users. Based on user influence obtained from skyline query, the users are ranked to find the most influential user. The process is conducted using top-k query according to [33]. This is because the algorithm does not depend on value function of data and is effective in determining the best object. Additionally, the algorithm can be processed in MapReduce framework. The following points are steps for top-k algorithm in MapReduce framework.

a. Data form skyline query was split according to data attribute (m) $\{S_1, S_2, S_m\}$. Value m in this study was 3.
b. Data m+1 $\{S_{m+1}\}$ is made, which the value is addition of three previous attributes.
c. In map process, data is ranked in each part according to ascending function and then process the ranking.
d. Map process output is a pair of user and ranking value (r) in each data that represents temporary key-value.
e. Combiner step is processed for the user pair and r form map process, which groups user-based-data.
f. The output from previous step is a pair of user and r with similar user. This output is used in reduce step.
g. In reduce step, sum-r is processed in each user group.
h. The output of reduce step is pair of user and sum-r.
i. The ranking is done according to ascending function and sum-r value. User with low sum-r is user influence on data.
j. Pair of user and sum-r is saved in accordance with value k.

3.   Results and Analysis
3.1. Data

In this work, a focus on Indonesian national figures were taken pleace. They are selected for the topic in data collection are Joko Widodo (jokowi), Jusuf Kalla, Ridwan Kamil (ridwan kamil) and Basuki Tjahaja Purnama (Ahok). These names are gained from identification process in January 2016. Based on these names, this study creates two groups of data containing opinion for the topic, corpus data and election.

First, corpus is data used as corpus to create model of sentiment classification. Data contain tweet collected in April 2016. The filtration is 1281 tweet consisting of 891 positive tweet, 289 negative tweet, and 101 neutral tweet. Second, selection data is used for identification and determination of user influence. The data are collected in April 2016. Identification of spam or bot is processed in data using Twitter follower-following ratio (TFF), which is useful to discard spam or bot users. TFF result is exhibited in Figure 2. Users identified as spam are users with TFF value approximately 0 or under black line. This stage results in 65.702 users with 551.513 tweet.
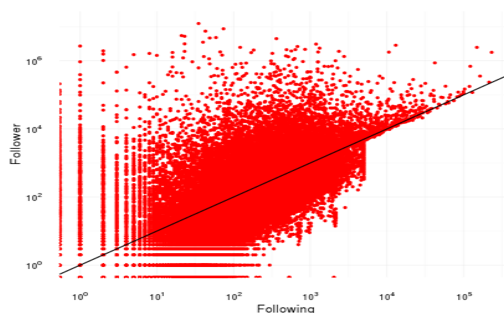

Figure 2. TFF ratio

3.2. Sentiment analysis model

Sentiment analysis with a setup of hybrid approach on corpus data shows accuracy 75.57% by applying a 10-fold cross validation. Thus, the model was applied for classification of selection data to determine sentiment weight of the user. Table 1 illustrated number of tweet and user in selection data with sentiment groups. In selection data, 11.326 tweet are tweet reply.

Table 1. Tweet and user in each sentiment group

	Positive	Neutral	Negative	Total
Tweet	182.932	184.289	184.302	551.523
Number of tweet user	35419	35629	35.752	65.702
Tweet reply	12.213	12.375	12.289	36.877
Number of tweet reply user	5.350	5.420	5.301	11.326

3.3. Characteristics value of user influence

BPMN in identifying user influence is presented in Figure 3. Initial identification is by analyzing data attribute of Twitter, calculating MT value, and determining sentiment weight of user. Based on data observation, Twitter data consists of 3 groups of attribute, attribute connecting with user activity, user profile, tweet content.
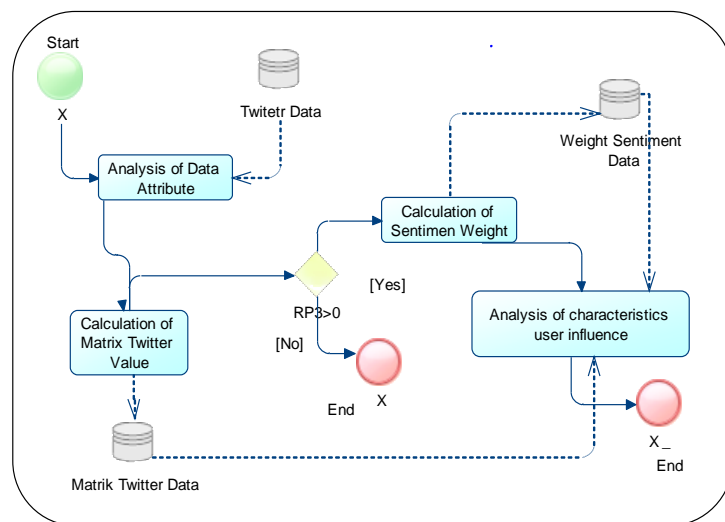

Figure 3. BPMN for identification of user influence

User activity is an attribute that connects with the user activity on Twitter. User profile is data attribute that contains user information related to user and its relation to other users including name, number of follower, following, etc. Tweet content describes information in tweet, such as user_mention, hastag, and symbols.

Table 2 is results of data attribute analysis that is usable for calculating MT value. The calculation is based on each user in selection data and attribute in Table 2. User that has ID RP3 is more than zero (RP3>0) is counted its sentiment weight using Equation 1. Data with MT value and sentiment weight are used to select value from each characteristic of user influence.

$$BS_u = \sum S_{u\text{-}r} \tag{1}$$

BS_u is sentiment weight u which is an value of reply sentiment ($S_{u\text{-}r}$) on tweet of user u in data. Sentiment 2 for positive sentiment reply, -2 for negative sentiment reply, and 0 for positive sentiment reply.

Tabel 2. ID matrik Twitter dan atribut data Twitter

ID	Atribut Data
F1	User:Follow ers_Count
F2	User:Follow ers_Count
F3	User: Friends_Count
F4	User: Friends_Count
F5	User:Screen_Name, Text
F6	User:Screen_Name, Text
M1	User:Screen_Name, Entities: User_Mentions
M2	User:Screen_Name, Entities: User_Mentions
M3	User:Screen_Name, Entities: User_Mentions
M4	User:Screen_Name, Entities: User_Mentions
OT1	Text
OT2	Url
OT3	Hashtag
RP1	User:Screen_Name, In_Reply_To_Screen_Name
RP2	User:Screen_Name, In_Reply_To_Screen_Name
RP3	User:Screen_Name, In_Reply_To_Screen_Name
RT1	Retw eeted_Status, User: Screen_Name
RT2	Retw eeted_Status, Retw eeted_Status:User:Screen_Name
RT3	Retw eeted_Status, Retw eeted_Status:User:Screen_Name
FT1	User:Screen_Name, User:Favorites_Count
FT2	User:Screen_Name, Favorited(True)
FT3	User:Screen_Name, User:Favorites_Count

3.3.1. External Communication

External communication indicates popularity. In Twitter, the popularity can be observed from F1 and F3 of MT, number of follower and following. User with the highest follower maybe not because of its tweet content quality, but it is socially active. Consequently, it has many friends and followers. Moreover, it may possible for user with the highest follower since it shares useful information, which is subsequently followed by other users. Therefore, value of external communication is calculated using equation 2 [34]. The user with external communication value approximately 1 is the best user based on this characteristic.

$$\text{FollowerRank=} \frac{F1}{F1 + F3} \tag{2}$$

3.3.2. Accessibility

Accessibility in Twitter based on MT as shown in Table 2 as seen from ID value F3, RP1, RT1, FT1, M2, appointed to following, reply, retweet, favorite/like, and mention of a user. User with high following activity shows preference for discussion with other users. A user is a following of other users, thus the user can reply, retweet, like and mention. Hence accessibility is gained from addition of each MT that represents user accessibility, calculated using formulation in equation 3.

$$\text{accessibility=} F3 + RP1 + RT1 + FT1 + M2 \tag{3}$$

3.3.3. Innovation

Innovation value in Twitter based on MT can be observed form OT1 and RT2, namely tweet and tweet that is re-tweet by other users in other users. In addition, innovation is observable from tweet sentiment given through reply from other users. Sentiment is represented by sentimen weight. Thus innovation value is obtained by adding value from each MT that represents innovation characteristic with sentiment weight. Innovation value is calculated using formulation in Equation 4.

$$\text{innovation=} OT1 + BS_u + RT2 \tag{4}$$

3.3.4. Implementation

Implementation compares between conventional approach and MapReduce. Conventional model is used to investigate efficiency and speedup MapReduce model. Conventional model is used in a single computer using programming model object oriented with aid of library Google guava. MapReduce uses library Hadoop in 3 computers (node). Table 3

shows comparison of execution time between conventional model and MapReduce in 3 nodes. The use of MapReduce enhances execution time (speed up) in counting characteristic value of user.

Tabel 3. Speed up for identification of user influence

	Conventional	MapReduce	Speed up
External communication	598,318 second	16,974 second	35,249
Aaccessibility	3.105,096 second	28,566 second	108,610
Innovation adaptation	468,139 second	22,973 second	20,378

Figure 4 shows graphic of time execution in MapReduce model framework with different nodes and data. Addition of node in MapReduce model shows shorter execution time and stagnant tendency in each variation of data number when data is more than 500.000 tweet. Data variations used are 100.000, 200.000, 300.000, 400.000, and more than 500.000.
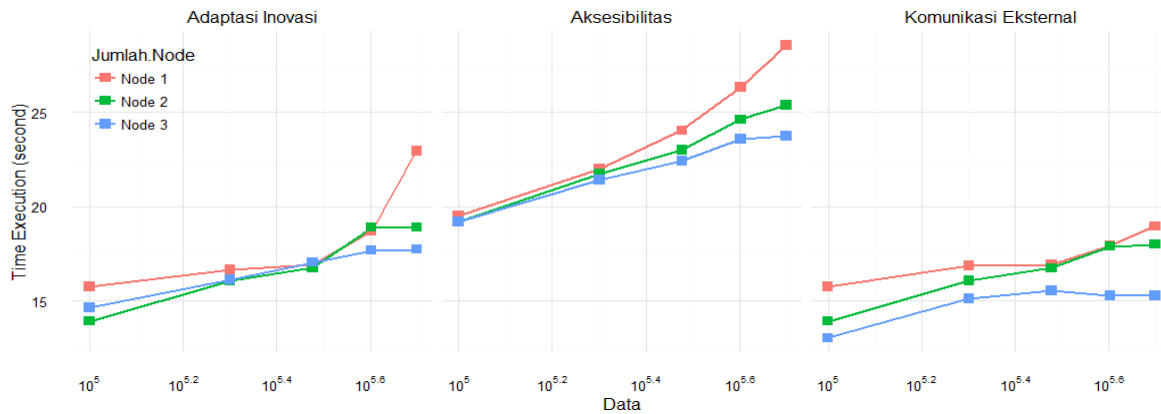

Figure 4. Comparison of execution time in different node number and data in each characteristic of user influence

Therefore, additional node in data results reflected consistency of time execution, which also suggested that MapReduce model enables to distribute time complexity to each involved node. Output of this stage is user data with value from each characteristic of user influence.

3.4. User influence

Data obtained from previous step consist of 3 attributes that represent characteristic number of user influence. Based on data, calculation is needed using algorithm SFS. Local skyline obtained are 186 users that are user influence in each part and 16 users for global skyline. Calculation of top-k query is required to determine the most influential user.

Figure 5 illustrates that bigger red dots or dots on the area are user influence in data. Figure 6 shows comparison execution time of conventional model and MapReduce in selection of user influence in skyline and top-k query. Data used are 100.000, 200.000, 300.000, 400.000, and more than 500.000 users. The k value used is 2, 4, 6, 8, and 10. Processing with skyline query shows shorter time execution and constant using MapReduce in each various node and data. However conventional model demonstrates exponential time execution in each different data.

In algorithm top-k with various k and node, execution time is not significantly different and constant either conventional model or MapReduce due to small k value. Nevertheless, the use of MapReduce model always has shorter execution time in various data. Table 4 is an example of user influence with value k=5. User influence by name "andreas_o_joeda" selected.
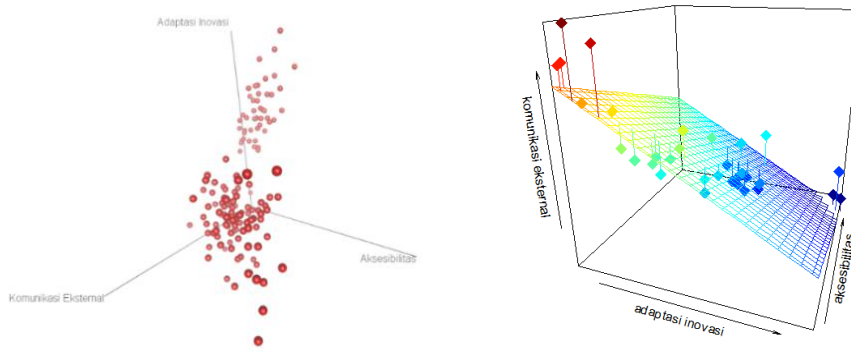
Figure 5. Illustration of user influence distribution and result of skyline query
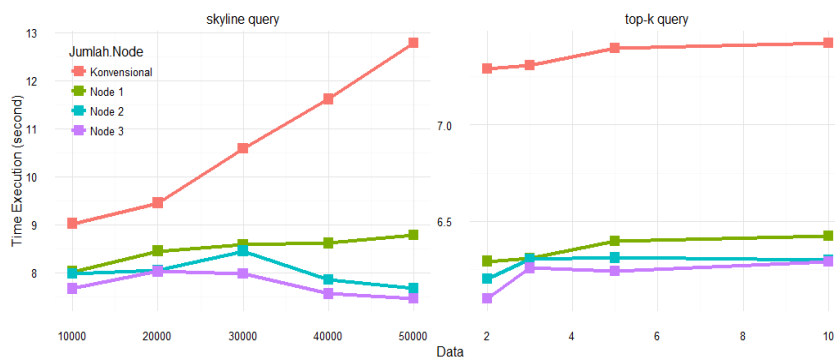

Figure 6. Comparison of execution time of conventional model and MapReduce on Skyline and top-k query

Application of skyline and top-k query in MapReduce framework in selecting user influence points out user that is able to influence other users in shorter time. However selected users do not discuss national figures-related topics corresponding to the issues in online media portals.

Table 4. Result of user influence determination with k=5

User	External Communication	Accessibility	Adaptation Innovation
andreas_o_joeda	8.293,21	941	349
ridw ankamil	4.028,25	233	245
beasisw aindo	344,18	176.842	56
kapanlagicom	400,99	3.097	789
metro_tv	256,032	2	43

4. Conclusion

Identification process of user influence showed that OL characteristics-external communication, accessibility, and innovation adaptation-are applicable for finding user influence in Twitter data. Based on these characteristics, the determination of user influence using by MT values and tweet content analysis achieved 16 user influences out of 65.702 users. In addition, identification process and user selection in MapReduce model demonstrates faster execution time in comparison with conventional model.

In case of future trand and investigation, the determination of user influence should be conducted in dynamic data style. Therefore selection process potentially address to current and real time position of user influence according to current issue of national figures.

References

[1] Maynard D, Funk A. *Automatic detection of political opinions in tweets*. in Proceedings of the 8th international conference on The Semantic Web (ESWC'11). 2012: 88-99;

[2] Jansen BJ, Zhang M, Sobel K, Chowdury A. 2009. Twitter power: Tweets as electronic word of mouth. JASIST 60(11):2169-2188. doi: 10.1002/asi.21149

[3] Kiss C, Bichler M. Identification of influencers-measuring influence in customer networks. *Decision Support Systems*. 2008; 46(1): 233-253

[4] Cha M, Haddadi H, Benevenuto F. *Gummadi PK Measuring User Influence in Twitter: The Million Follower Fallacy*. ICWSM. 2010; 10: 10-17.

[5] Riquelme F, González-Cantergiani P. Measuring user influence on Twitter: A survey. IPM. 2015; 53: 949-975

[6] Huang L, Xiong Y. Evaluation of microblog users' influence based on pagerank and users behavior analysis. AIT. 2013; 3(2): 34-40.

[7] Septiandri AA, Purwarianti A. Identifikasi Opinion Leader pada Twitter dengan Teknik Pembelajaran Mesin. Konferensi Nasional Informatika. 2013

[8] Raamakirtinan S, Livingston LJ. Identifying Influential Users in Facebook-A Sentiment Based Approach. *IJST*. 2016: 9(10): 1-9

[9] Zaman A, Siddique MA, Morimoto Y. *Selecting Key Person of Social Network Using Skyline Query in MapReduce Framework*. Proceedings of the Third International Symposium on Computing and Networking (CANDAR,15). 2015: 213-219

[10] Djatna T, Morimoto Y. Reviving Dominated Points in Skyline Query. *IJCSNS*. 2009; 9(1):347

[11] Tong W, Xuecheng Y. *Electronic word of mouth in online social networks*. ICCSNA. 2010; 2: 249-253.

[12] Jiang Z, Dai W, Cao X, Chen L, Zheng K, Sidibe A. Research on Community Detection Algorithm Based on the UIR-Q. *TELKOMNIKA*. 2016: 14(2): 725-733.

[13] Rogers EM. Diffusion of innovations. New York (AS): The Free Press.1995.

[14] Dean J, Ghemawat, S. *MapReduce: simplified data processing on large clusters*. in Osdi'04 proceedings of the 6th conference on symposium on operating systems design and implementation.2004:107-113.

[15] Budi I, Bressan S, Wahyudi G, Hasibuan ZA, Nazief BA. *Named entity recognition for the Indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach*. Proceedings of the 8th international conference on Discovery Science (DS'05). 2005: 57-69.

[16] Rozi FI, Pramono HS, Dahlan AE. Implementasi opinion mining (analisis sentimen) untuk ekstraksi data opini nasional pada perguruan tinggi. *Jurnal ECCIS*. 2008; 6(1):37-43.

[17] Medhata W, Hassan A, Korashy H. Analisis sentimen algorithms and applications: A survey. ASEJ. 2014; 5: 1093-1113.

[18] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. CS224N Project Report. 2009; 1:12.

[19] Wahyudin I, Djatna T, Kusuma WA. Cluster Analysis for SME Risk Analysis Documents Based on Pillar K-Means. TELKOMNIKA. 2016; 14(2): 674-683.

[20] Adityawan E. Analisis sentimen dengan klasifikasi naïve bayes pada pesan Twitter menggunakan data seimbang. Thesis. Bogor: Graduate; 2014.

[21] Hu M, Liu B. 2004. Opinion Lexicon. http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

[22] Vilares D, Thelwall M, Alonso MA. The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. JIS. 2015; 41(6): 799-813.

[23] Guerini M, Gatti L, Turchi M. *Sentiment analysis: How to derive prior polarities from SentiWordNet*. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013:1259-1269.

[24] Wilson T, Wiebe J, Hoffmann P. 2005. MPQA Subjectivity. http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[25] Witten IH, Frank E, Hall MA. Data Mining: Practical machine learning tools and techniques. Fourth Edition. Morgan Kaufmann. 2005.

[26] Illecker M. Real-time Twitter Sentimen Classification base on Apache Storm. Tesis. Innsbruck: Postgraduate Universitas Innsbruck; 2019.

[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. 2002; 10:79-86.

[28] Borzsonyi S, Kossmann D, Stocker K. The skyline operator. Proceedings of the 17th International Conference on Data Engineering. 2001:421-430.

[29] Chomicki J, Godfrey P, Gryz J, Liang D. *Skyline with presorting. in Data Engineering*. 2003 Proceedings. 19th International Conference on. 2003: 717-719.

[30] Kossmann D, Ramsak F, Rost S. *Shooting stars in the sky: An online algorithm for Skyline queries*. In VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases. 2002: 275-286.

[31] Papadias D, Tao Y, Fu G, Seeger B. Progressive skyline computation in database systems. TODS. 2005; 30: 41-82.

[32] Zhang B, Zhou S, Guan J. *Adapting Skyline Computation to the MapReduce Framework: Algorithms and Experiments*. Proceeding DASFAA'11 Proceedings of the 16th international conference on Database systems for advanced applications. 2011: 403-414.

[33] Siddique MA, Morimoto Y. *Efficient Selection of Various k-objects for a keyword Query based on MapReduce Skyline Algorithm*. DNIS 2014 Proceedings of the 9th International Workshop on Databases in Networked Information Systems. 2014: 40-52.

[34] Nagmoti, R, Teredesai, A, De Cock, M. Ranking approaches for microblog search. In Web Intelligence and Intelligent Agent Technology (WI-IAT). 2010; 1: 153-157.