Journal of Engineering and Applied Sciences 12 (Special Issue 3): 6406-6411, 2017
ISSN: 1816-949X

Utilization of Social Media for Consumer Behavior Clustering using Text Mining Method

Harwati, Agus Mansur and Adnan Karunia
Department of Industrial Engineering, Islamic University of Indonesia, Yogyakarta, Indonesia

Abstarct: Social network has grown rapidly in line with the development of technology. Twitter is one of social media that has worldwide. It is widely used among individual user as well as companies or organizations to communicate with their consumer. Consumer's tweet can be utilized by companies to determine consumer behavior in response to their product or service. This method is more efficient than mining voice of by distributing questionnaires. Comments written by customers is usually tend to original and not contrived. The objective of this research is to map the consumer behaviors based on the opinion in Twitter. Smartphone iPhone 5 as a trending topic smartphone in Twitter is taken as the object for this research. Clustering text mining techniques are used to settle the problem. Two level clustering is done to get the mapping of consumer behavior. In the first step, RapidMiner is used to cluster more 200 comments in to 8 groups. Second step is called by profiling groups. It obtains three clusters by analyzing the similarity characteristics and the meaning of the comments contained in each group. The resulting final cluster are cluster of 29.20% positive comments, cluster of 42% negative comments and a cluster of other comments by 28.80%. From this result, it can be concluded that based on their comments in the social media Twitter, the unsatisfied consumers is greater than others.

Key words: Twitter, consumer, behavior, clustering, text mining, RapidMiner

INTRODUCTION

Along with the development of internet technology, social media also experiences rapid enhancement. Social media website provides easiness in accessing the information and interactions without considering the distance and time. This development brings significant impacts on everybody's life style. Also, improvements in information technology have made it easy for company to collect enormous amount of customer transaction data (Sohrabi and Khanlari, 2010). Initially, social media designated to entertainment, then become the marketing phenomenon due to its enormous advantages in business (Kirtis and Karahan, 2011). Twitter is one of the world widely social media websites. In April 2013, it reaches 500 million users. The average tweets reach 58 million tweet (www.statisticbrain.com). Recently, there are more companies employ social media as tools such as Facebook and Twitter to provide various of services and to interact with customers (He *et al*., 2013).

Products from apple which is iPhone 5 also well known worldwide. iPhone 5 is a product from apple that mostly discussed in Twitter. iPhone 5 ever be the trending topics in Twitter almost in every country all over the world (www.en.trending-topic.com). Every kind of discussion about iPhone 5 are deliberated in Twitter. Based on existing surveys, three biggest retailers for iPhone 5 in Yogyakarta, only employ Twitter as means of promotion and only 1 out of 3 uses it to communicate with customers. They use Twitter for promotion and inform customers about exisiting promotions. Recently, there are no further analysis about customer's comments about iPhone 5. Comments or tweets that discussed about iPhone 5 issues are considered as huge data that could be used to identify customer's behavior.

Qiu *et al*. (2012) in their research, showed that tweet that contains linguistic hint legally could represent someone's characteristics. From this statement, Twitter can be used to recognize customer's behavior. In fiercer global business competition, it is important to identify potential customers. Company not only expands their market for new customers but it is important to maintain potential and existing customers. Customer's behavior is dynamic, easy to be changed as time. Hence, an industry should be able to move and change as the requirement of its customers.

Customer organizational citizenship behavior or often abbreviated as customer OCBs is defined as individual behavior out of direct role of company but yet, its participation and capabilities will impact the company

Corresponding Author: Harwati, Department of Industrial Engineering, Islamic University of Indonesia, Yogyakarta, Indonesia

(Bove *et al*., 2009). In customer OCBs, there are several groups of costumer's behavior that differentiated based on the stated words. There are eight dimension of costumer's behavior in customer OCBs which are: positive statement, advice for service improvement, affecting other customer, costumer's voice, good treatment on service facility, demonstrate good relationship, flexibility and participation in company's activities (Bove *et al*., 2009). By gathering these costumer's statements on products, company could identify its costumer's behavior. A lot of comments could be processed by using one of data mining methods which is clustering technique to obtain group of costumer's behavior. Later, it could be implemented as consideration on strategy improvement of marketing management. Clustering is a component of exploratory data analysis and is useful for generating hypothesis about data (Mohammadkhanlo and Bashiri, 2013).

Clustering of a set forms a partition of its elements chosen to minimize some measure of dissimilarity between members of the same cluster (Pakhira, 2014). There are several methods in clustering technique, one of them is K-means cluster. In this technique, the number of classes will be initially determined. These traditional clustering approaches generate partitions and every pattern is associated with one and only one cluster (Sohrabi and Khanlari, 2010). Hence, this research will employ eight clusters to classify comments derived from social media of Twitter. Text Mining is automatic deriving to find hidden patterns form text sources. Text mining has been used in several research to gaining customer knowledge in several areas (Liau and Tan, 2014). Text mining is a flexible method that could be applied in the wide scope of database studies and problems, particularly in the form of text. One of the problems in text mining is calculation in recognizing hidden pattern in text. Considering that the data derived from Twitter are in the form of text, hence, text mining will be utilized in accomplishing this research. By using clustering method associated with text mining technique, costumer behavior grouping based on comments in Twitter will be conducted in this research. Furthermore, the classes of customer's behavior will be clearly identified and can be used in determining better marketing strategy. Based on above background, hence the title of this research can be resumed as "The utilization of social media Twitter as means of customer's behavior clustering by using text mining method approach".

MATERIALS AND METHODS

The purpose of this research is to determine and identify customer's behavior for iPhone 5 based on their comments in social media of Twitter by using clustering method based on text mining, so, it can be later identified its advantages to marketing strategy. Data for this research are primary that derived from social media of Twitter. Tweets data from Twitter user accounts will be extracted by employing extraction website which is www.zapier.com. This research also applies Google Docs as storage media. Technically, zapier.com will conduct data extraction based on given keywords. Later on, automatically, the extracted data will be stored in Google Docs. For that intention, users should have accounts in both above tools. Keyword for data extraction is iPhone 5.

Yang and Su (2012) conducted research about customer's behavior clustering by using SVM method (Support Vector Machine). SVM was used due to its superiority in identifying data pattern, especially for small sample, nonlinear and high dimension problems. In this research, clustering was conducted to obtain more advantages which were identifying potential requirement from costumers and to expand the market by learning the customer's behavior. This research conducted extraction on customer's behavior and later used it as the input for proposed model. It proved that SVM method was effective and could be processed easily, yet needed further research.

Other research was conducted by Lakshminarayan *et al*. (2005). In their research an analysis on customer's experience over website's service. It used visitor's comments as objects that were expressed in website. This research used text mining method as tool to process text comments. Furthermore, clustering method was executed on costumer's comments data for easy understanding. Later by using classification techniques which are SVM and NBC (Naive Bayes Classifier) to categorize texts. The purpose of the research was to demonstrate that text mining could be employed to extract valuable information from text data.

Guo and Wang (2010) studied costumer's behavior in field of communication. This research analyzed costumer's behavior clustering by using fuzzy C-means clustering method. This research used incoming and outgoing call duration as data that later clustered by fuzzy C-means clustering. Basic of clustering was ARPU (Average Revenue Per User) to classify the customer's behavior. Therefore, there were three clusters for each ARPU.

Luo *et al*. (2009) executed research on text documents clustering based on document's familiarity or often called as neighbors. The objective of this research was to propose two different methods of neighbors which are link in k-means and bisecting K-means to cluster the documents. This research suggested that link function

Table 1: Cleaning data process

Before	After
iPhone 5 (Metalico)/iPhone 5/iPhone 5	Deleted
RT @AppIe Officiall: the first	Deleted
1oo people that retweets	
this & amp	
RT"@jenniferJLM: The iPhone 5 has the wors battery life! I have to take my charger everywhere"	The iPhone 5 has the worst battery life! I have to take my charger everywhere
"@gorgeouschassy: I hate iPhone 5 chargers!!!"!!!!!	I hate iPhone 5 chargers!
So happy to have my iPhone	So happpy to have my iPhone 5
I luv iPhone 5 earpiece	I love iPhone 5 earpiece

provided global perspective in evaluating closeness between two documents by using document's similarity or neighbors.

Li *et al.* (2008) conducted research about text documents clustering based on word order frequency. This research developed recent text clustering methods which are Clustering based on Frequent Word Sequence (CFWS) and Clustering based on Frequent Word Meaning Sequences (CFWMS). This research resulted that, two above proposed methods could cluster the documents better and capable to cluster unique documents with high dimension or applied sensitive language context.

RESULTS AND DISCUSSION

Data preparation process: After establishing tweet data with iPhone 5 as keyword then, the data will be prepared or experiencing preprocessing before being processed. The preprocessing is conducted to avoid invalid data from clustering process. Data cleaning is used as stage of preprocessing. It is applied to let software identifies texts properly before process it maximally.

Data cleaning for tweets, covers:

- Only includes tweets in English
- Selects tweets that only comment on products
- Ignores mentions and re-tweet symbol without changing the words or intention of a tweet
- Omits the excessive punctuation marks on tweets
- Omits or convert the symbols with appropriate words associated with symbol's expression
- Justifies inappropriate words with correct English
- Delete the tweets that contain words which contain untrue meaning

Processing of cleaning data is shown as following Table 1. In data processing, data are clustered in eight clusters. Each of eight clusters has words similarity characteristics. Determination of cluster's division in to eights is based on research carried out by Bove *et al.* (2009) stated that customer's behavior based on their statements are divided into 8 groups which are: positive statement, advice for service improvement, affecting other customer, costumer's voice, good treatment on service facility, demonstrate good relationship, flexibility and participation in company's activities. Yet, they cannot be implemented to see customer's behavior cluster on iPhone 5, since the formed clusters are based on the words similarity, instead of the word's meaning.

Therefore, all eights clusters are interpreted manually for easy understanding to be later analyzed for its costumer's behavior.

Clustering used software RapidMiner to help the process became more efficient. The interface of clustering process using RapidMiner is shown in Fig. 1 and 2.

Profilization: Cluster 1 with percentage of 3.60% shows that comments are grouped based on word similarity of camera. Costumers provide feedback on camera that become one of specifications of iPhone 5. Most of them feel satisfied and pleased with camera that planted in iPhone 5. Cluster 2 with percentage of 10%, the comments are grouped based on battery word. From several comments, costumers complain about the battery performance which is run out quickly.

Cluster 3 with percentage of 12.40% is grouped based on word of fuck. In this cluster, costumers express their disappointment on several features of iPhone 5 by using the word of fuck. Most of disappointments involve the performance of charger.

Cluster 4 with percentage of 7.60% with similarity characteristics of word hate. Regarding to the content of cluster, costumers dislike iPhone 5 charger that easily broken. Next cluster which is cluster 5, the percentage is 12.00 that has similarity with the word of shit. Similar with previous clusters, this cluster represents customer's frustration to charger by using the word of shit.

Cluster 6 has the biggest percentage which is 28.80%, since it contains several dominant keywords. They are charger and phone. For charger keyword, averagely costumers feel disappoint while for phone keyword, several costumers compare iPhone 5 with other products. Cluster 7 with the percentage of 18.80% with comments cluster characteristics that use the word of want. In this cluster, comments show their enthusiasm towards iPhone 5. More customers want to have iPhone 5. Cluster 8 has percentage of 6.80%. In this cluster, the word that becomes its characteristics is love. Customers represent their satisfaction towards iPhone 5 by using

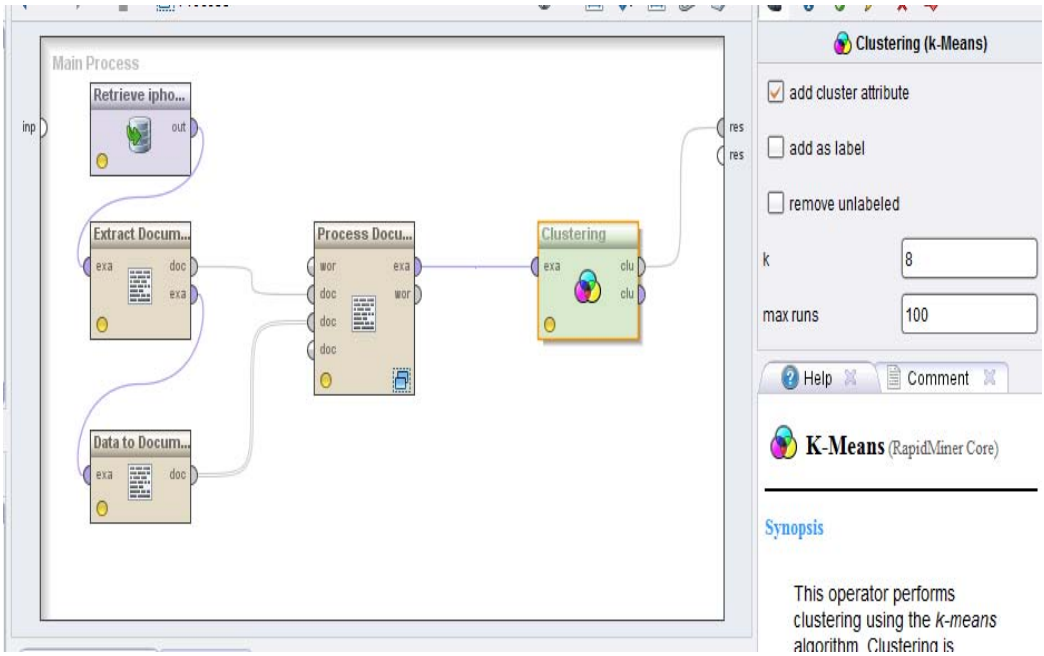

Fig. 1: Clustering process using RapidMiner

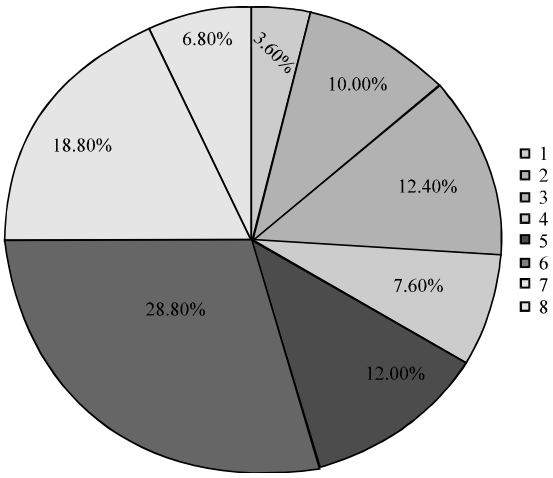

Fig. 2: Percentage for each cluster from eight clusters

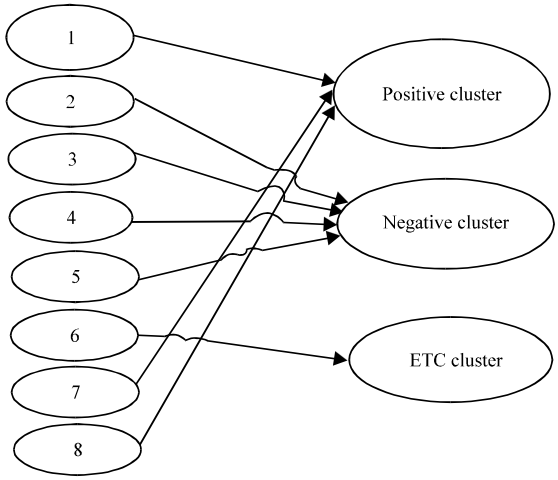

Fig. 3: First clustering in to 8 cluster

designated word. All eight clusters then are classified into three clusters by considering the meaning of words on how the customers express their comments by Twitter. Further classification in to three cluster is executed, so the data can be more understandable by considering the meaning contains in each comment. This classification is based on the research by Bove *et al.* (2009) stated that there are eight clusters for each customer's behavior based on the comments. Three clusters are explained as positive comments cluster, negative comments cluster and other comments cluster. Positive comments cluster covers comments that demonstrate positive behavior over iPhone 5. Negative comments cluster represents the opposite. While other comments cluster covers other comments that based on weight calculation are classified in to the same cluster (Fig. 3 and 4) (Moshiri *et al*., 2003).

Positive comments cluster is composed from cluster 1,7 and 8 from previous clusters. From whole data, positive comments hold 29.20%. Positive comments

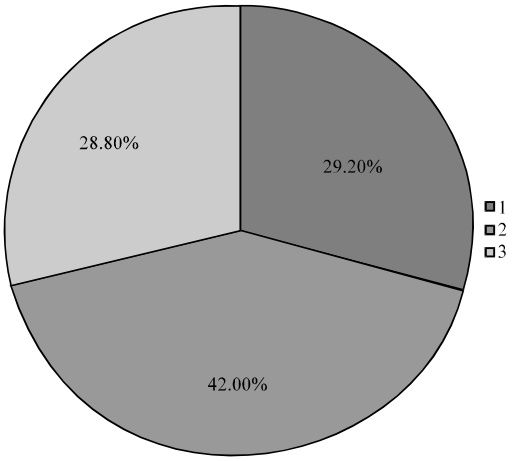

Fig. 4: Cluster grouping and percentage of three clusters; second level of clustering

cluster is formed by considering the existing comments. In cluster 1, comments are grouped based on the word of camera. Even though camera is not an adjective but from the content of comments, they represent satisfaction on the presence of amera in iPhone 5. Hence, cluster 1 is included in positive cluster. Cluster 7 is formed based on the similarity of the word want that composes every sentences. The word of want suggests customer's wish to have iPhone 5, surely this can be considered as customer's positive behavior for aspiring and requiring iPhone 5. Cluster 8 is grouped based on the similarity word of love that composed in every comment. The word of love represents happiness or appreciation in using iPhone 5.

Negative comments cluster is composed from cluster 2-5 from previous clusters. It covers 42.00% from overall data. Negative cluster is formed by considering the negative comments inside it. Cluster 2 has characteristics with the word of battery. This cluster is included to negative comments cluster since, the majority of comments demonstrated their disappointment to the performance of battery. Based on existing comments, it was stated that battery for iPhone 5 runs out quickly, this was the main source of customer's complaints. In cluster 3-5, the grouping was based on similarity of words fuck, hate, shit and suck. From their meaning, it can be classified as negative comments. Most of comments complained the charger of iPhone 5 that can be easily broken. Other comments complained about battery of iPhone 5 that run out quickly. Cluster 3 that was grouped based on similarity of word fuck contains several comments that had different meaning with the real meaning of word fuck. This is caused by the execution of clustering that only considers the result of weighting. Hence, the meaning of words in those comments will be ignored.

Other comments cluster is derived from cluster 6 of previous clustering. This cluster holds 28.80%. There are several dominants words which are charger and phone. In this cluster, the comments are mixed. The assortment possibly caused by the clustering ignores the meaning of words, so the negative and positive comments might be combined. This is caused by the similarity of the component's weight so the k-means cluster detects it as similar comments or documents that finally, merged in to single cluster. There are various comments in this cluster, there are comparisons with other products, comments on issues of product's charger and battery and other comments that express happiness for having and using iPhone 5.

CONCLUSION

From the data processing by using clustering methods based on text mining, tweets or comments in social media of Twitter could be grouped in to three clusters of customer's behavior, positive comments cluster with 29.20% that covers word's indicator of want, camera and love. Negative comments cluster that hold 42.00% with word's indicator of battery, fuck, hate and shit. Other comments cluster with percentage of 28.80 with no specific main words.

REFERENCES

Bove, L.L., S.J. Pervan, S.E. Beatty and E. Shiu, 2009. Service worker role in encouraging customer organizational citizenship behaviors. J. Bus. Res., 62: 698-705.

Guo, Z. and F. Wang, 2010. Telecommunications User Behaviors Analysis Based on Fuzzy C-Means Clustering. In: Future Generation Information Technology, Kim, T.H., Y.H. Lee, B.H. Kang and D. Slezak (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-17568-8, pp: 585-591.

He, W., S. Zha and L. Li, 2013. Social media competitive analysis and text mining: A case study in the pizza industry. Int. J. Inf. Manage., 33: 464-472.

Kirtis, A.K. and F. Karahan, 2011. To be or not to be in social media arena as the most cost-efficient marketing strategy after the global recession. Procedia Soc. Behav. Sci., 24: 260-268.

Lakshminarayan, C., Q. Yu and A. Benson, 2005. Improving Customer Experience via Text Mining. In: Databases in Networked Information Systems, Bhalla, S. (Ed.). Springer, Berlin, Germany, ISBN:978-3-540-25361-7, pp: 288-299.

Li, Y., S.M. Chung and J.D. Holt, 2008. Text document clustering based on frequent word meaning sequences. Data Knowledge Eng., 64: 381-404.

Liau, Y.B. and P.P. Tan, 2014. Gaining customer knowledge in low cost airlines through text mining. Ind. Manage. Data Syst., 144: 1344-1359.

Luo, C., Y. Li and S.M. Chung, 2009. Text document clustering based on neighbors. Data Knowledge Eng., 68: 1271-1288.

Mohammadkhanloo, M. and M. Bashiri, 2013. A clustering based location-allocation problem considering transportation costs and statistical properties research note. Int. J. Eng. Trans. C. Aspects, 26: 597-604.

Moshiri, B., P. Eslambolchi and R. Hoseinnezhad, 2003. Fuzzy clustering approach using data fusion theory and its application to automatic isolated word recognition. Int. J. Eng. Trans. B., 16: 329-336.

Pakhira, M.K., 2014. A fast k-means algorithm using cluster shifting to produce compact and separate clusters. Int. J. Eng. Trans. A. Basics, 28: 35-43.

Qiu, L., H. Lin, J. Ramsay and F. Yang, 2012. You are what you tweet: Personality expression and perception on Twitter. J. Res. Personality, 46: 710-718.

Sohrabi, B. and A. Khanlari, 2010. Targeting customers: A fuzzy classification approach. Int. J. Eng. Trans. A. Basics, 23: 323-335.

Yang, Z. and X. Su, 2012. Customer behavior clustering using SVM. Phys. Procedia, 33: 1489-1496.