Clustering on Twitter: case study Twitter account of higher education institution in Indonesia

Almed Hamzah* and *Ahmad Fathan* Hidayatullah

Department of Informatics, Universitas Islam Indonesia, Jalan Kaliurang KM 14,5 Yogyakarta, Indonesia

Abstract. Recently, higher education institutions have been using Twitter as one of tools to enhance their communication network. This paper aims to cluster Twitter data retrieved from the official Twitter account of higher education institutions in Indonesia. We expect to obtain a valuable information from the tweet posted. Furthermore, we use Twitter's hashtag as a basis of clustering. We collect data from n=10 institutions that have an official account on Twitter. The Affinity Propagation algorithm was employed to perform the clustering task. According to the clustering results, we conclude that higher education in Indonesia mostly utilize Twitter to post general information, news, agenda, announcement, information to the new students, and achievement.

1 Introduction

The popularity of social media is increasing most rapidly over the last decade. Social media has become a trend in people's daily life over the last decade. Social media has been utilized by many individual and institutions to share any information about various topics [1].

In the academic context, social media is the one of the tool that could help academic activities. Among the social networking sites, Twitter is the one of social media that has been widely used by academic institutions such as higher education. Previous research has reported that Twitter is the one of the most well-known social media in the world and has significant influence in the academic world compared to other social media including Facebook [2]. Therefore, there are some benefits that can be obtained by higher education from using Twitter for academic purposes.

This research aims to extract information from Twitter data which posted by higher education in Indonesia based on Twitter hashtags. In addition, Furthermore, the clustering task would be performed to identify the topic group the hashtags. Therefore, the result of the clustering task could illustrate the meaningful topic and what kind of topics that posted by higher education in Indonesia.

The rest of the paper is organized as follows. Section 2 will describe about review result of related works that have been done by other researcher. Section 3 will explain about methods that has been used in this study. Section 4 will deliver the research result and followed by relevant discussions. Section 5 will end this paper by concluding remarks.

2 Related Works

In this part, the related works which has been conducted before are discussed. Rosa, et al. [3] have investigated about tweet classification into some categories. The approach of the research inspired from Google News service provided by Google. They used LDA and K-Means to cluster the tweet data. Moreover, they also proposed new method to complement the lack of Twitter in annotating the topic using hashtag. The method to classify the tweet are unsupervised and supervised method. The unsupervised method has revealed that the tweet classification conducted based on the language or word similarity not based on the discussed topic. On the other hand, the supervised method has worked better on the short text and high noise.

Tang, et al. [4] have reported that there are two issues on tweet clustering process. Firstly, the issue about data sparsity caused by the short message on Twitter which limited no more than 140 characters. Secondly, the issue about the language ambiguity which frequently used by people in tweet messages. Accordingly, they proposed the concept from Wikipedia in representing tweet into vector concept.

Tripathy, et al. [5] proposed the clustering technique based on word frequency and Wikipedia topic taxonomy to find the discussed topic in the tweet. The research has revealed that the proposed algorithm has given better result than the algorithm which only involve word frequency.

Kim, et al. [6] has suggested the method called Core-Topic-based-Clustering (CTC) to extract meaningful topics then cluster the tweet based on the topic. As comparison, the research also conducted the clustering task using K-Means method. The experiments showed that CTC method can efficiently extract the meaningful topic. In addition, the CTC algorithm has better performance than K-Means.

The use of social media in the context of the

* Corresponding author: almed.hamzah@uii.ac.id

institutions has not been done. This is most visible from publications that there are currently only discusses the use of social media by individuals, although some publications discussed in the context of the academic world. This is similar to that expressed by [2]. However, a few publications are trying to reveal the role of social media at the university level.

Based on studies conducted by [2], there are two main interests that encourage the use of social networking media by the university, i.e. marketing activities and delivering information about the university to the students. These findings are based on the perception of the university; however, the findings are not supported by empirical data. Other studies revealed similar things, even supported by empirical data. This research was conducted by [7] found that, on Twitter, universities use it for marketing and public relations (50%) and communicating with students (50%).

3 Data Retrieval

The data in this research were retrieved from the Twitter account of top 10 universities and colleges in Indonesia based on the rank from the minister of research, technology and higher education in 2015. Before retrieving tweet data, we identified the official Twitter account of those top 10 higher education in Indonesia. The tweet datasets were retrieved using Twitter API v1.1. To access the API, we utilized Tweepy library in Python. Data retrieval was carried out by using GET statuses/user_timeline provided by Twitter. GET statuses/user_timeline method provides a set of tweets from a particular user with screen_name or user_id parameters.

The downloaded Twitter messages are the tweets which are not protected by the owner of the account. Furthermore, the tweet data obtained using the above method is similar to the tweet data viewed by the user profile on Twitter. However, this method can only access approximately 3240 latest tweets from a particular user timeline. We have obtained 31351 tweets with various amount of Tweets for each account being observed.

4 Topic Clustering Based on Hashtags using Affinity Propagation

In this section, the topic clustering task based on hashtags would be explained. This step aims to group the tweets based on their similarity of hashtag-based topics. In Twitter, users use hashtag (#) to represent the subject of their posts. Hence, we can extract the topic of the messages by identifying their hashtag only. Rosa et al. [3] stated that tweet clustering based on hashtag has given better result than content based clustering.

4.1 Affinity Propagation

The chosen algorithm to cluster the hashtags is Affinity Propagation (AP). This algorithm clusters the data by sending messages among the data points until convergence is reached [8]. Before running the algorithm, it is not necessary specify the number of clusters since it works by iteratively refining a randomly-chosen initial set of exemplars, which are recognized as those most representative of other samples.

The algorithm works through a number of iterations until the data has reached the convergence condition. Moreover, there are two messages passing categories in each iteration, responsibility and availability.

The responsibility *r(i,k)* measures how well-suited *k* is to serve as the exemplar for point *i*. Furthermore, the responsibility of a sample *k* to be the exemplar of *i* is updated by the function:

$$r(i,k) \leftarrow s(i,k) - max\left[a(i,\acute{k}) + s(i,\acute{k}) \forall\, \acute{k} \neq k\right] \quad (1)$$

The availability *a(i,k)* shows how appropriate the point *i* to be picked in sample *k* as its exemplar. The availability calculation is based on the function:

$$\mathrm{a}(i,k) \leftarrow min\left[0, r(k,\acute{k}) + \sum_{i\, s.t.\, i \notin \{i,k\}} r(\acute{\imath},k)\right] \quad (2)$$

Initially, all values for *r* and *a* are set to zero. After that, the calculation will run iteratively until reach the convergence condition.

4.2 Hashtags Analysis

From the Twitter dataset, we extract all of the hashtags. The extraction process of hashtags from Twitter was conducted by identifying words or phrases preceded by the hash (#) symbol. Overall, there are 12813 hashtags and 1825 unique hashtags from total of 31351 tweets. These hashtags will become raw data for clustering task. Figure 1 shows the top 20 most frequent hashtags in our dataset.

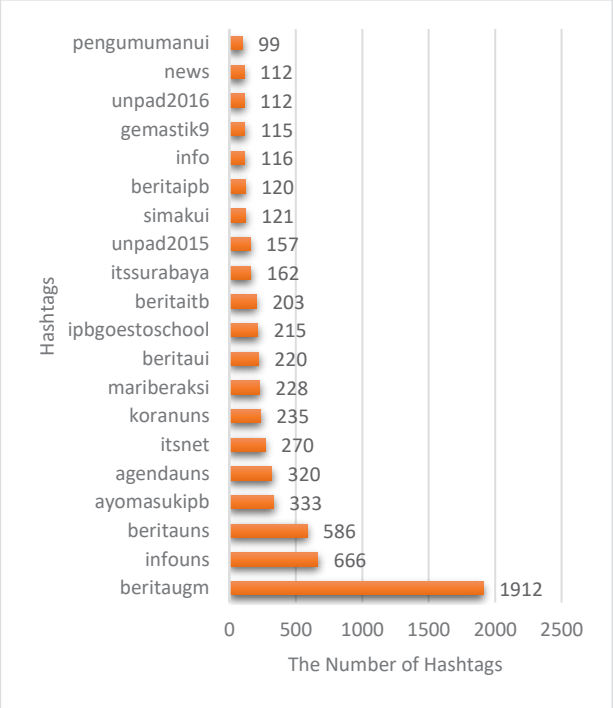

Fig 1. Top 20 most frequent hashtags

The top 20 most frequent hashtags illustrate the most topic that are frequently posted by the Twitter account of higher education in Indonesia. At a glance, the information topic posted are talking about news or general information about the college.

According to the figure, it can be seen that half of top 20 represents about news or information, for instance *beritaugm, infouns, beritauns, koranuns, beritaui, beritaitb, beritaipb, info, news,* and *pengumumanui.* The other hashtags indicate the agenda or activity such as *agendauns, ipbgoestoschool, simakui,* and *gemastik9*. The hashtags *itsnet, itssurabaya, unpad2015,* and *unpad2016* could be identified as the general information about the university itself. The rest of the hashtags such as *ayomasukipb* and *mariberaksi* illustrates persuasion to do specific action.

4.3 Clustering Result

The hashtags clustering was performed by using the Affinity Propagation algorithm. Before the clustering process, we compute the similarity matrix by multiplying negative and Levenshtein distance. In addition, this research works with precomputed distance measure by assigning *"precomputed"* value to affinity parameter. The convergence iteration is 10 which means the number of iterations with no change in the number of estimated clusters that stops the convergence after 10 iterations. Maximum number of iterations are 100 and damping factor is 0.5.

According to the clustering process, the estimated number of clusters is 274 as can be seen in figure 2. However, the silhouette coefficient of this result is -0.358. This low silhouette value indicates that the clustering configuration may have too many or too few clusters.

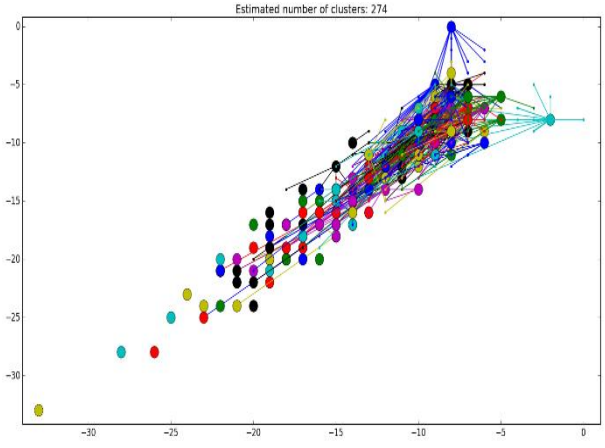

Fig 2. Clustering plot

Table 1 shows the top ten cluster results with the hashtags in the clusters. We tried to label each of the cluster by looking the most frequent hashtags in the cluster.

Table 1. Top ten hashtags clustering results

Clusters	Hashtags	Topics
Cluster – 1	38thuns, cdcuns, expouns, fotouns, ikauns, info, infocdugm17, infogizi, infolomba, infonh, infopkm, infougm, infoui, infouns, internship, lingkungan, lnfouns,	Info
Cluster – 2	agenda, agenda2, agendabandung, agendafh, agendaipb, agendanh, agendaugm, agendauns, agendaunsuns,	Agenda
Cluster – 3	koranuns, civitasuns, congratssss, garansindo, gelorarun, gowesuns, hope4uns, kabaruns, karyapenuhmakna, kerjahumas, klipinguns, kontakuns, koranuns, koranunsuns, korauns, kotasolo	News about UNS (Universitas Negeri Sebelas Maret)
Cluster – 4	beritafema, beritafh, beritafik, beritafkui, beritaipb, beritaitb, beritaskpm, beritaugm, beritaui, beritauns, berituns, berkasub	News
Cluster – 5	econews , enews, ieo17, ieyc, junewish, nes, new, news, newyear, niche, nots, now, runners, sdgs, ubnews, unsnews	News
Cluster – 6	pengumuman, pengumumandiib, pengumumanfik, pengumumanui, pengumumanuns, penundaanwisuda,	Announcement
Cluster – 7	ub, ub51, ub53, ubdi, ubstlh, ubtv, ui, uktub, uns, uts	Specific information about colleges
Cluster – 8	camabaub2013sila, camabaub2014, camabaub2103, mabaub2013	New Students of Universitas Brawijaya
Cluster – 9	award, budaya, calonjuara, charoen, eksyarjuara, itsjuara, iysca, jakarta, jgtc39, juara, juaraan, juarai, jumatberkah, mobiljuara	Achievement
Cluster – 10	drpmupdate, fkhmdelegates, funnyquote, liveupdate, puprgtcw, quote, unstoppable, unsupdate, update, updates	Information update and quotes

It can be seen that the hashtags in the cluster-1 are dominated by the word *"info"* which means that the topic in cluster-1 is general information about the college. The hashtags in cluster-2 denote the agenda of the higher education. Cluster-3 is more specific talking about the news regarding UNS (Universitas Negeri Sebelas Maret). Cluster-4 and 5 represent the news topic. The hashtags in

cluster-6 are about the announcement. Cluster-7 shows the specific information about the colleges. In cluster-8, all the hashtags represent the information of new students of Universitas Brawijaya. Cluster-9 is talking about the achievement of the colleges. Cluster-10 shows the information update and quotes.

5 Discussion

In this section, we will discuss further about the clustering results. According to the hashtags analysis in the previous section, it is mentioned that a half of the top 20 hashtags depicts the news and general information regarding the colleges. The top 10 clustering results show that cluster-1, cluster-3, cluster-4, cluster-5, cluster-7, and cluster-10 also represent about news and general information on the higher education. Therefore, it can be concluded that most of higher education in Indonesia mainly utilizes Twitter to post general information and news.

Furthermore, there are more specific topic that represented by the cluster-2, cluster-6, cluster-8, and cluster-9. The topic of cluster-2 has the same topic with the previous hashtags analysis which illustrate about the agenda. From the cluster 2, 6, 8, and 9, there are some important topics that frequently posted via Twitter such as agenda, announcement, information to the new students, and achievement.

6 Concluding Remarks

Twitter is among the alternatives for higher education to held an effective communication with their stakeholders. It can be a means to share various information regarding to academic activity. Moreover, it allows the user to create specific topic by using Twitter's hashtag which in turn will increase reader's awareness about particular information. In this study, we use the hashtags to cluster the information that being posted by universities on their Twitter's account.

The clustering method we used in this research is Affinity Propagation Algorithm. According to the clustering results, we conclude that higher education in Indonesia mostly utilize Twitter to post general information, news, agenda, announcement, information to the new students, and achievement.

Overall, from the Twitter messages posted by the higher education in Indonesia could give the illustration about what kind of information that frequently discussed. The information posted could be give beneficial for the internal people in the colleges and external people outside the colleges.

The authors would like to thank Department of Informatics Universitas Islam Indonesia to fund this research in the scheme of Internal Research Grant.

References

1. J. K. Sinclaire and C. E. Vogus, "Adoption of social networking sites: An exploratory adaptive structuration perspective for global organizations," *Inf. Technol. Manag.*, vol. 12, no. 4, pp. 293–314, 2011.
2. C. H. F. Davis III, R. Deil-Amen, C. Rios-Aguilar, and M. S. G. Canche, "Social Media in Higher Education: A literature review and research directions." 2012.
3. K. Dela Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical clustering of tweets," *Proc. ACM SIGIR SWSM*, 2011.
4. G. Tang, Y. Xia, W. Wang, R. Lau, and F. Zheng, "Clustering tweets using Wikipedia concepts." in *LREC*, 2014, pp. 2262–2267.
5. R. M. Tripathy, S. Sharma, S. Joshi, S. Mehta, and A. Bagchi, "Theme Based Clustering of Tweets," in *Proceedings of the 1st IKDD Conference on Data Sciences*, 2014, pp. 1–5.
6. S. Kim, S. Jeon, J. Kim, Y.-H. Park, and H. Yu, "Finding core topics: Topic extraction with clustering on tweet," in *Cloud and Green Computing (CGC), 2012 Second International Conference on*, 2012, pp. 777–782.
7. R. Reuben, "The use of social media in higher education for marketing and communications: A guide for professionals in higher education." 2008. Retrieved from http://doteduguru.com/wp-content/uploads/2008/08/social-media-inhigher-education.pdf
8. B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science (80-.),* vol. 315, no. 5814, pp. 972–976, 2007.