Int. J. Advance Soft Compu. Appl, Vol. 11, No. 3, November 2019
ISSN 2074-8523

Spam Detection on Profile and Social Media Network using Principal Component Analysis (PCA) and K-means Clustering

Samuel Ady Sanjaya and Kridanto Surendro
School of Electro and Informatics Engineering,
Institut Teknologi Bandung, Bandung, Indonesia
e-mail: samuelady@gmail.com, surendro@gmail.com

Abstract

Social media as a means of communicating in cyberspace continues to grow both from the number of users, utilization, and the resulting impact. Existing social media ecosystems are influenced by the influence of public figures, trending topics, even spam, and spammers. Detection of spam accounts that have been done mostly using the method of classification or supervised learning. This will be a problem if the data is new and the supervised model is not updated it will increase the possibility of false detection. Based on the problem, this study will use Principal Component Analysis (PCA) and K-means clustering with Mahalanobis distance as a method to detect a collection of users who have similar properties to determine spam. This study uses 150 thousand twitter data with 15 thousand account data that described as graph data. The result, we find that error detection in the classification method to find spam is a class that made only two: spam and non-spam. Though in addition there are still other classes that have the characteristics of spam when it is not. In this paper, we defined the clusters on to 5 clusters: normal, news account and public activist, foreign account, public figure, and spam.

Keywords: *K-means, Principal Component Analysis (PCA), Social Media, Social Network Analysis, Spam.*

1 Introduction

Web evolution has introduced new applications that facilitate interactive information exchange, user-based content creation, and remote collaboration. Contrary to static web pages common in the past, where users can only access information, today's web apps allow users to communicate, upload and modify content [6]. Traditional media, for example, newspapers, magazines, TV, and radio can only provide the diffusion of one-way information, as the audience can read or listen, but can not share their opinions about a subject. Instead, social media provides a two-way information diffusion that allows users to share their information with other users, access information and communicate. Social media is social interaction among people where they create, share, or exchange information and ideas in virtual communities and networks [13]. Social media has also been defined as a group of Internet-based applications built on the foundation of ideology and Web 2.0 technology and which enables the creation and exchange of user-generated content [11]. Communication and sharing that are common in social media are to comment on someone's status on social networks or to rank videos on video sharing sites, or more complex as movie recommendations [6]. However, social media is not often used as a tool to spread lies, spam, or do things that are not fair [18, 21].

In this study, we propose an approach to detect spammers with unsupervised models that do not require data train which requires a long process. The proposed model can also be applied to another new dataset since it is not specifically created for one dataset. Use of features from profiles and social network analysis makes the dimensions bigger, so it needs to do feature reduction. In the end, the combination of PCA as the dimensional reduction method and k-means clustering as an unsupervised model is the right choice to use in this research. In this study, we also evaluate that not only spammers have outstanding feature values but there are other clusters that have value specific to describe the cluster. The rest of this paper is organized as follows. In Section 2, we briefly introduce several related works on social spammer detection and spam text detection. In Section 3, we discuss the research methodology, and the optimization algorithm to solve the model of our approach. In Section 4, we report the experimental result of son a Twitter dataset. In Section 5 we conclude this paper.

2 Related Work

In some studies, user behavior or unnatural content is described as an outlier [5]. An outlier is a term used to describe abnormalities, aberrations or anomalies in data mining and statistical literature [1]. The appearance of outliers is largely unexpected, but outliers often contain useful information about the abnormal characteristics of systems and entities that affect the process of data generation. A feature approach that can be used to detect users who have outlier characteristics is the profile, content, time and graph [11]. Research conducted by Shenepoor is to develop a graph-based framework that is used to detect spam based on a language of content and user behavior [21]. Spam detection with content features and user profiles using incremental learning is used to detect hashtag-based, profile and content spam with ever-increasing data [20, 17, 24].

The approach used in the previous study only used a subset of several approaches, so it does not represent objects that are classified as outliers as a whole. The methods used in previous studies mostly use supervised learning or classification methods that require labeled data. The classification method has a high degree of accuracy but lacks flexibility and requires time-consuming data labeling processes [8]. Based on this background, with the growing social media data required unsupervised learning methods to identify outliers more quickly and map the population data as a whole. The research will also use a variety of approach features to describe the overall outlier. Using many features will affect the increasing dimension of features that can reduce the accuracy of the classification model. The feature selection method is used for clustering method optimization by identifying feature relevance, feature independence, and feature weight counting [9]. The clustering method to be used is k-means, by looking at the characteristics of each cluster with abnormal centroids to be identified as outliers. With this research is expected to be developed a model for outlier detection based on profiles and user content on the scope of time and role in social media networks.

3 Methodology

The methodology is performed starting from twitter data collection, data preparation, and clustering. Data preparation is used to get what features will be used in spam detection. Network analysis is performed as part of the feature extraction to get a profile centrality measure in the network. normalization is used as a scaled-back for the data to be properly visualized. The next stage is the use of PCA for weighting features from social media data. Weighted data will be selected to continue in the next stage of clustering. The core of this research is to perform a PCA analysis to get feature rankings that have a strong relationship. After doing dimension reduction with PCA next is to clustering to divide data in accordance with the similarity between its value. After getting the cluster, it will identify which clusters fall into the spam category.

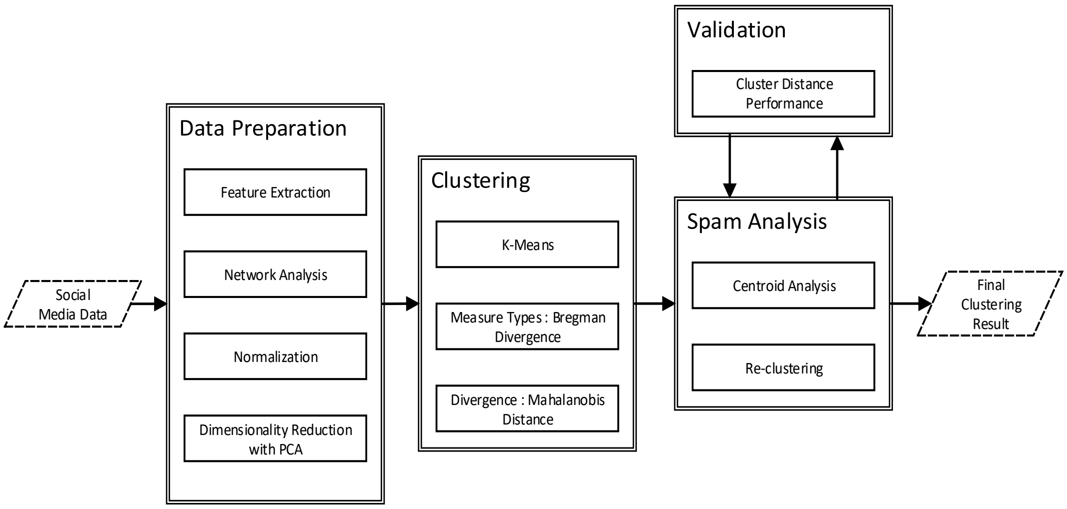

Fig. 1: Research Methodology for Spam Detection

For details of each stage will be explained in the description of the research methodology below from data collection to spammers identification data with PCA and K-means clustering.

3.1 Data Collection

The collection of data from social media is the first step in conducting social media analysis which will be used for outlier detection (see Figure 2). Social media data is taken from Twitter which provides the API to access tweets uploaded by users based on keyword, hashtag or username [6]. Twitter was chosen because it allows more open data retrieval through the API compared to other social media such as Facebook which is limited to public groups [21]. Intake of Twitter data is divided into two parts, namely twitter data based on keyword search and twitter data based on specific users. Twitter data retrieval is done for 14 days using the same keywords to get the sample data that represents the condition of social media and interest in the keyword.

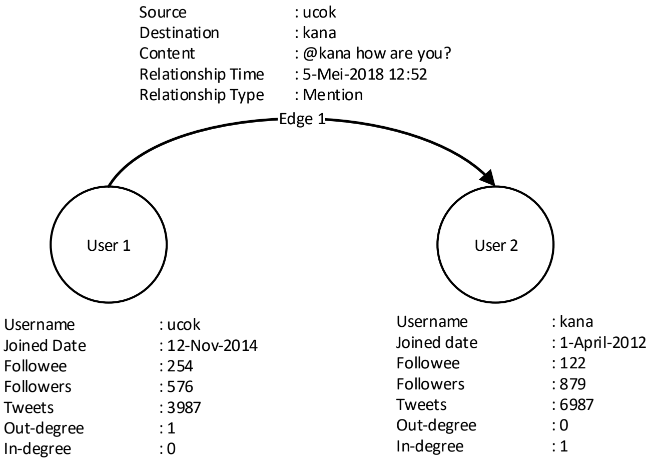

Fig 2: Example of social media graph data, the relation of one user and another user

Twitter provides data retrieval in the form of graphs that define users as nodes and content as edges as in Figure 2. Each node represents a user profile attribute that includes the number of followers, followed, tweets, the date the account was created and personal preferences. Each edge represents the metadata of an uploaded content that includes the date of uploading the content, the type of communication, the hashtag used and other linked users.

3.2 Feature Extraction

The feature extraction stage is a pre-process data stage to identify features that have the potential to describe an object [10]. Social media data extracted from the API is dirty data that can be optimized to get more in-depth information. Social media data is modeled in graphs to find relationships between users and measure the value of user roles in a network. The role of users in the network is known as the term centrality measure or measurement of user activity based on the edges contained in the network [22]. Other feature extractions are counting the number of symbols in a username, a number of mentions and URLs in a content, comparison between followed and followers, account age and reciprocated relationship ratios. Centrality measure or measurement of centrality is the method used to determine the person or object that affects the network [23].

3.2.1 Degree centrality

Degree centrality has two main components: in-degree and out-degree. Out-degree is the decision of the node to interact with other nodes within the network. In-degree is not a decision of the origin node, but it is a received response from another node that leads to the origin node. The degree centrality approach is used to identify the central user by counting the number of incoming interactions and outgoing interactions from a single node.

3.2.2 Closeness centrality

The centrality of nodes in this approach is defined as a node whose sum of vertex spacing with other vertices has a small value. The less closeness centrality the greater its proximity to the other node.

3.2.3 Betweenness centrality

Betweenness centrality measures the number of vertices used as "bridges" in the shortest path between the other nodes. A node will have a high centrality betweenness value when the probability of occurrence is high in determining the shortest distance between nodes.

3.2.4 Eigenvector centrality

Eigenvector centrality also commonly called Eigen centrality is the development of degree centrality. In the in-degree centrality, the degree of centrality is high when a node becomes the destination of another node through an edge. In in-degree centrality, the weight of all incoming nodes is computed while the approach is not. The weighted value of an incoming node is not the same because there are several factors that affect such as the relevance and support of the other node.

3.2.5 Reciprocated Ratio

he reciprocated ratio is a measure of the possibility of vertices to be related to each other in directed tissue analysis. This is a quantitative measure to assess the connectivity of a node with another vertex through the linking edge. Self-loop is ignored and in the non-directional network, this reciprocity cannot be defined.

3.3 Normalization

Social media data has a very high variation so that the distribution distance in each attribute is very large. Data that has a high variation will not be optimal to be processed especially in the clustering process because it causes the distance between the clusters to be uncontrolled. Normalization becomes the solution to give a new scale to each attribute so that the variations in each attribute are more controlled. Adjusting the scale values for each attribute will make it easier for analysis and comparison. One method of normalization is the feature scaling. Feature scaling is used to convert all values into ranges 0 and 1, also known as unity-based normalization.

3.4 Dimensionally Reduction with PCA

Large data can be classified because there are many features that reflect the entity. However, many features do not apply to results due to irrelevant or redundant features [10]. In addition, the many features processed in a classification model can slow down performance. For that, we need the selection of features to eliminate irrelevant features. Peng et al. [16] proposed a feature selection method that could use mutual information, correlation or distance/similarity scores to select features. The goal is to determine the feature relevance of the redundancy compared to other features selected.

The problem that often arises is "curse of dimensionality" where the machine difficult to calculate data with very much input [14]. One of the most common ways to handle this is to reduce data input while retaining the information contained within it. PCA can decrease the minimum dimension by maintaining the information contained in it. PCA is a linear transformation commonly used in data compression so that its dimensions become simpler. PCA is a useful statistical technique in various fields such as classification, data compression, unsupervised learning and more [22].

3.5 K-Means Clustering

Large volumes of unstructured data require data clustering techniques to find implicit patterns of data and information. Clustering as an analytical tool that aims to categorize the object into categories, namely the relationship between objects to a maximum including the same category [3]. K-means clustering method will divide the data group in accordance with the similarity between attributes [15]. The number of groups in k-means can be determined according to the value entered. Below in Figure 3 is a k-means algorithm flow for clustering against a set of data. With clustering, the data does not need to be labeled first so it is possible to directly process new data with large volumes without going through the process of labeling data.

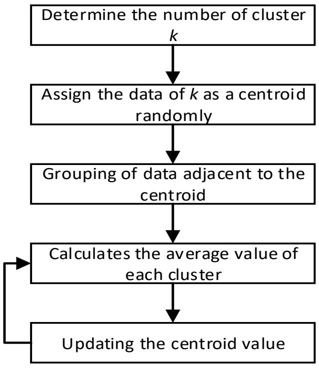
Determine the number of cluster *k*
Assign the data of *k* as a centroid randomly
Grouping of data adjacent to the centroid
Calculates the average value of each cluster
Updating the centroid value

Fig 3. Clustering using K-Means

There are several distance measurements between clusters called links that can be used for the merging of clusters. Mahalanobis Distance is a generalization of Euclidean Distance that is devoted to multivariable data calculations that have high variations [4].

3.6 Cluster Distance Performance for Validation

Cluster Distance Performance operators take these centroid cluster models and group sets as inputs and find models based on cluster centroids. There are two performance measures of an existing cluster, namely the average distance and the Davies Bouldin Index. Two performance measures are supported: Average cluster distance and Davies-Bouldin index. The average distance of the cluster is calculated by the average distance between the center and each member. Davies Bouldin is an algorithm that produces clusters with low intra-cluster distances and high inter-cluster distances will have a low Davies-Bouldin index, a grouping algorithm that produces a good collection of clusters with Davies-Bouldin index [7].

3.7 Spammer Identification

The last stage is to identify which clusters are indicated according to the characteristics of each cluster. Clusters that have the most characteristics as spammers are categorized as spammers, but if only just one or two characteristics do not. This is often a mistake in detecting where to find one of the characteristics directly defined as spam when it is not. Table 1 is derived from a collection of previous studies that have defined the characteristics of spammers [12]. This set of characteristics serve as the basis for determining spammers more fully, not just some characters.

Each cluster that has a feature value as in the table above is suspected as spam. Feature values are taken from the centroid value of each cluster that represents the hallmark of each cluster.

Table 1. Spammer characteristic based on previous research [12, 2]

No	Feature	Category	Value (Spam)
1	Number of Followers	Profile	Low
2	Number of Followed	Profile	High
3	Followed-Follower Ratio	Profile	High
4	Symbol on Username	Profile	High
5	Default Profile Picture	Profile	Yes
6	Default Profile	Profile	Yes
7	Verified Profile	Profile	No
8	URL on Profile	Profile	Yes
9	Account Age	Profile	Low
10	Favorite Count	Profile	Low
11	Geo-location Enabled	Profile	No
12	In Degree Centrality	Network	Low
13	Out Degree Centrality	Network	High
14	Out-In Degree Centrality	Network	High
15	Betweenness Centrality	Network	High
16	Closeness Centrality	Network	Low
17	Eigenvector Centrality	Network	Low
18	Page Rank	Network	Low
19	Reciprocated Ratio	Network	Low
20	Mention Count	Content	High
21	Hashtag Count	Content	High
22	URL in Content	Content	High
23	Average Account Activity	Content	High
24	Relation Count	Content	Mention
25	Media on Tweet	Content	Yes
26	Content Favorite Count	Content	Low
27	Retweet Count	Content	Low

4 Results, Analysis, and Discussions

4.1 Data Collection and Feature Extraction

The data collected is 15.455 Twitter user profiles from about 150.000 tweets. Analysis of data in the form of a graph is a feature extraction step to find the centrality of accounts in a network. centrality becomes important because it shows the role and interest of users in a topic. Centrality is rarely used because of its dynamic and changing nature over time. In this study, centrality analysis is used because it is considered to have a relationship with the character of users who are categorized as spam. Figure 4 is one of the visualization betweenness centrality that put "Jokowi" as the center of a network because the keyword used is "Jokowi" as an Indonesian President. From the picture, it can be seen that the node "Jokowi" is often passed by another node, or in other words "Jokowi" is often mentioned by other account causing big betweenness centrality.

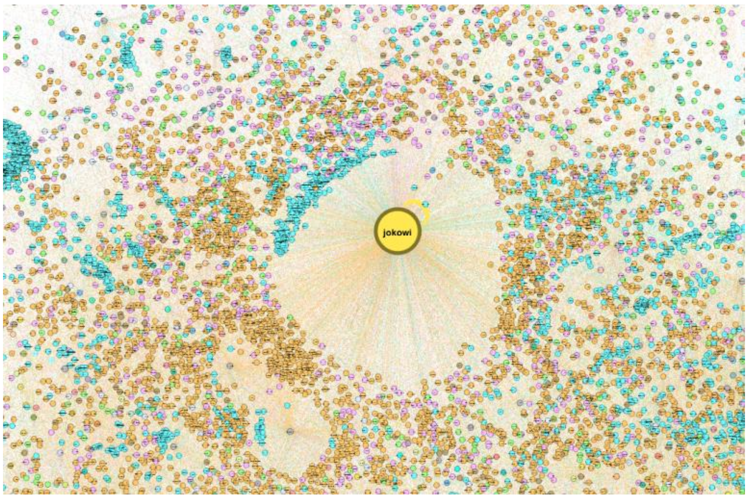

Fig 4. Betweenness centrality visualization as a part of feature extraction

After performing the centrality analysis, the next is to dig deeper into the features that are likely to be a prominent factor in cluster determination. Based on a collection of previous research and personal experiments, there are several features that are made into ratios, namely followed/followers ratio (FFR) and out/in-degree ratio (OIR). Account age expressed in week scale to simplify cluster. For exact features like true/false for verified, default profile image, and default profile are used as numeric attributes to be counted in clustering.

4.2 Dimensionally Reduction with PCA

PCA is used to weighting attributes that are the basis for choosing attributes for clustering. The result of this research is the clustering of social media profile data in accordance with the attribute equation owned by each profile. Prior to clustering, to optimize the clustering results used PCA as a method to reduce dimensions. PCA is used for feature ranking based on its correlation with other attributes. From the ranking of features with PCA, we get the weight of each feature. High-weight features will be selected for clustering with k-means. Based on the data that has been obtained, the results of feature ratings with PCA can be seen in Figure 5.

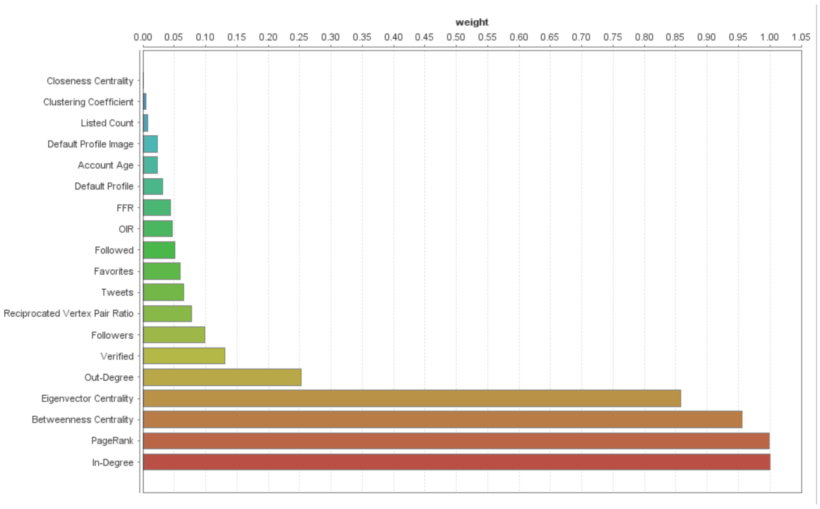

Fig 5. Weighting with PCA for dimensionally reduction

Based on the weights we've got, we use 7 features to use for clustering. Although only 7 attributes are selected, will still be discussed an overall feature that describes a profile object.

4.3 K-means clustering

In K-means clustering, it is determined the number k to divide the cluster of k. the number of k clusters used is 20, many of these clusters are used to group social media profile data into specific clusters. The distance measurement used is the common Mahalanobis Distance used to measure the distance between data in many dimensions. Mahalanobis distance is considered more able to cope with data with many dimensions compared with Euclidean distance which is only optimal on 2-dimensional data measurement. From the initial clustering results found several clusters that have prominent characteristics as spam. The emergence of Jokowi as outlier data may affect the visibility of other data. Figure 6 localization has been done so that every feature will be spread on a scale of 0 to 1.

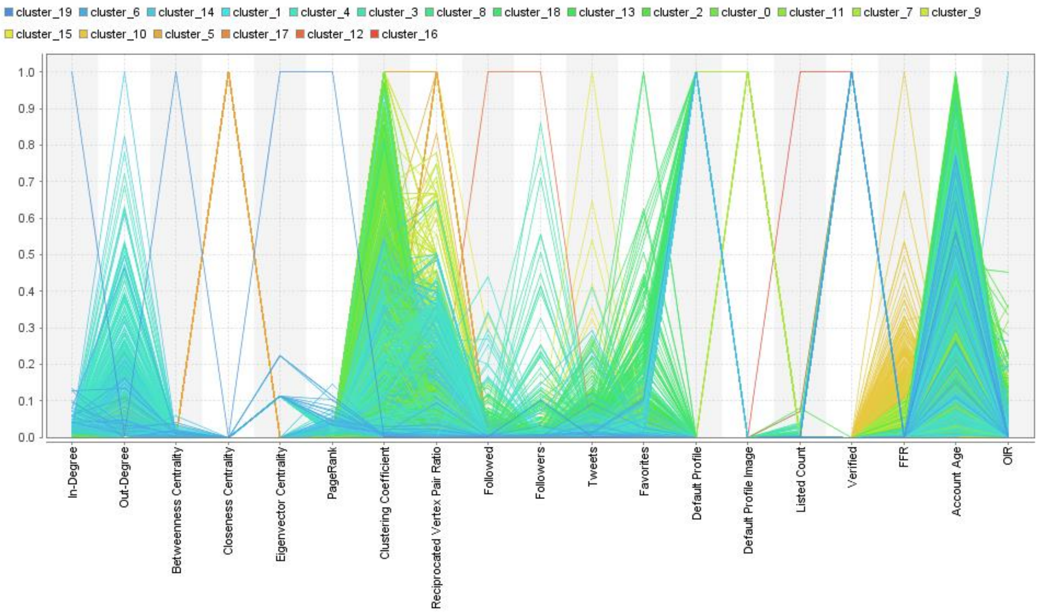

Fig 6. 20 cluster with local normalization per-feature

4.4 Cluster Distance Measure

The Davies-Bouldin Index (DBI) is an internal evaluation scheme, which is used to validate how well the groupings have been carried out using the quantity and features in the dataset. This condition limits the index so that it is defined as symmetrical and not negative. Because the way it is defined, as a function of the ratio in clustering the lower value cluster will mean that grouping is better. This is the average similarity between each cluster and the most similar, the average of all clusters. In table 2 we tried to use many *k* of clusters, and the application of PCA. The results in table 2, show that the use of PCA and more *k* in clustering produce better cluster distance average, even though the DB index is not as good as small clusters but the results are quite satisfying compared to not using a PCA.

DB index and average cluster distance which the value is the lowest is a good measure of the number of clusters data could be ideally classified into. Average results from the average distance of the cluster and DBI with *k=20* and the application of PCA show good number so that the next process for cluster simplification can be done.

Table 2. Average cluster distance and Davies Bouldin index

		No-PCA		PCA	
No	*k* cluster	Avg. Centroid Dist.	DB Index	Avg. Centroid Dist.	DB Index
1	2	0.661	1.108	0.395	1.065
2	10	0.217	1.868	0.185	1.255
3	15	0.173	2.151	0.122	2.311
4	18	0.163	2.317	0.089	2.073
5	**20**	0.164	2.737	**0.084**	**1.864**

4.5 Spammer Detection

Of the 20 clusters that have been defined before, cluster simplification will be done by grouping clusters that have a centroid value that is defined as a normal profile into one. So in this study, we grouped from 20 clusters into 5 clusters that have the same properties. The profiles that have unique characteristics that describe the value of spam as in table 1, will still be used as a different cluster. Here is the result of the centroid tables of each cluster. The simplification of 20 clusters into 5 clusters has been labeled as in table 3, with the centroid value of each feature used.

Table 3. Average centroid value of the clustering result after clustering simplification (5 clusters)

Features	Spam	News and Active User	Verified Public Figure	Normal	Unexpected Account
Verified	0.0026	0.4226	0.6667	0.0117	1.0000
Tweets	41112.0644	728463.6647	32056.3667	21042.1857	25417.2745
Reciprocated Ratio	0.0252	0.0880	0.0064	0.2048	0.0000
PageRank	0.8230	4.7038	412.8382	2.1112	1.6415
Out-Degree	5.5571	12.8009	9.6667	13.9765	0.0000
OIR	0.9866	1.2215	0.0196	0.7417	0.0000
Listed Count	41.7007	1085.4031	1979.2667	31.0970	1178796.9804
In-Degree	4.1716	34.0399	3548.1667	14.8198	5.7059
Number of Followers	30473.7889	694034.8408	5293244.8667	27249.5923	49122377.4314
Number of Followed	4105.1612	57245.1166	605.3000	713.5415	211789.5294
Favorites	39676.5753	8577.8005	4969.6000	4403.2538	2033.7451
FFR	18.7011	0.5290	0.0095	2.6979	0.0022
Eigenvector Centrality	0.0000	0.0000	0.0050	0.0002	0.0000
Default Profile Image	0.2952	0.0000	0.0000	0.0095	0.0000
Default Profile	0.8941	0.3107	0.1333	0.5526	0.0196
Clustering Coefficient	0.2449	0.2396	0.0056	0.1847	0.1994
Closeness Centrality	0.0000	0.0417	0.0000	0.2386	0.0000
Betweenness Centrality	11702.9598	206925.6201	75283160.1430	41375.1048	88677.4974
Account Age	24.4170	75.7467	74.5319	57.9229	103.2708

From the 5 clusters, there is one cluster that is the unexpected account where this account contains accounts coming from outside the keyword that accidentally entered so as to have an influence in the analysis. For easy viewing from table 4, we visualize it in figure 7 with the graph radar. From the graph can be seen the characteristics of each cluster according to their centroid value.

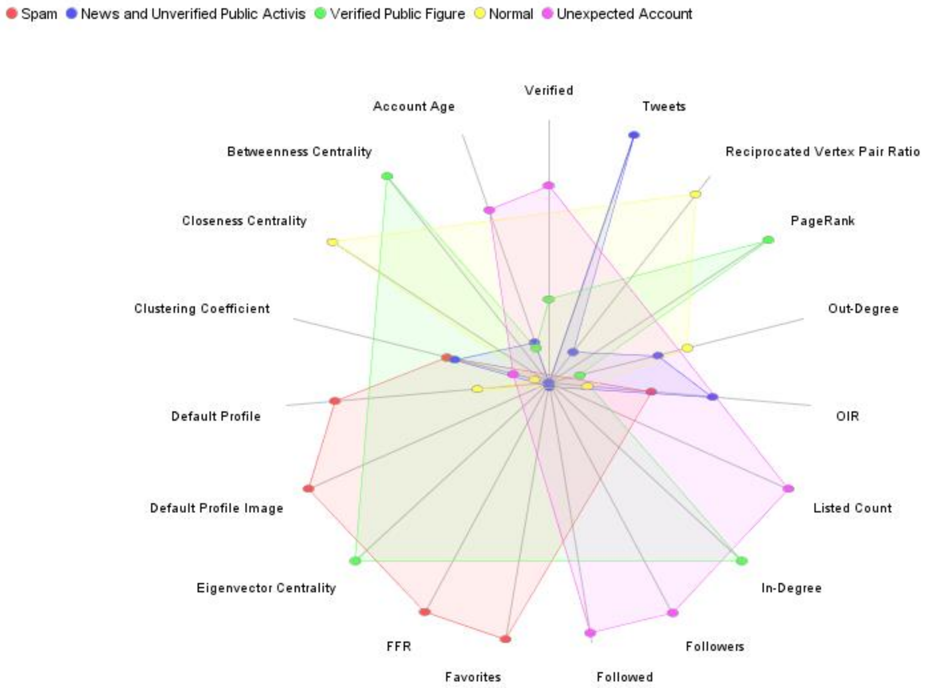

Fig 7. Visualization of centroid value

The detail of user count every final cluster can be seen in Table 4. From the 5 clusters, there is one cluster that is an unexpected account where this account contains accounts coming from outside the keyword that accidentally entered so as to have an influence in the analysis. Cluster with a label of unexpected account contains users from abroad, some of whom are public figures so they have verified accounts and have a large follower’s base. From the visualization, it is clear that each cluster has its own characteristics, where spammers have a tendency to have a much more followed than followers and the age of his account is relatively shorter.

Table 4. User count every cluster

Cluster Name	User Count
Figure Public	410
Foreign Accounts	19
News and Active User	3,374
Normal	5,422
Spammer	6,230

Spammers have a small reciprocated vertex pair ratio as well because the communications are run in one direction, in contrast to normal users who have high reciprocated vertex pair ratio where two-way communication occurs. In spammers networks, they have high OIR because spammers attack more often than getting a response from other users. The public activist and news account have a high out-degree but also in-degree, so this cluster cannot be defined as a spammer. The news and public activist cluster consist mostly of a variety of news accounts that do every day to update the news obtained but cannot be said as spam because it does not attack other users. For verified public figures clusters, has a large centrality in the network because it becomes the center of other users.

5 Conclusion

This study produced a model for spammer detection based on user profiles and content on the scope of time and role in social media networks. The result found that error detection in the classification method for finding spam is a class that created only 2: spam and non-spam. Though in addition there are still other classes that have the characteristics of spam when it is not. Public figures appear as accounts that are indicated as spam because they have a high centrality betweenness, but other attributes don't support it like the account status verified by Twitter. While news accounts and active users are indicated as spam because they have a higher out / in degree state rate ratio than normal user accounts.

So the clustering method with k-means and PCA can identify clusters that have more specific characteristics and spam characteristics, not just divided into two as in supervised learning. News accounts and active users still cannot be categorized as spam because they still have a reciprocal interaction ratio and an odd account age. Spam accounts identified in this study combine several aspects of the characteristics of spam from previous research and then make a comprehensive analysis, not only as a feature. The 5 clusters established there were 6,230 accounts detected as spam, 19 including in the unexpected accounts, 3,374 accounts included in active users and news accounts, 410 as public figure accounts and 5,422 including normal accounts. This research was conducted to develop the identification of spam and other classes that indicated spam but actually not by utilizing profiles and social media networks. Subsequent research may be able to develop this research by participating in analyzing the content in the form of links, videos or images linked to a post.

Acknowledgment

The authors would like to thank the Institut Teknologi Bandung (ITB) for their support in Research and Development, Information System Lab for the inspiration in making this study a success.

References

[1] Aggarwal, C. C. (2016). *Outlier analysis*. *Outlier Analysis* (Vol. 9781461463). https://doi.org/10.1007/978-1-4614-6396-2

[2] Al-Zoubi, A. M., Alqatawna, J., & Faris, H. (2017). Spam profile detection in social networks based on public features. *2017 8th International Conference on Information and Communication Systems (ICICS)*, 130–135. https://doi.org/10.1109/IACS.2017.7921959

[3] Alsayat, A. (2016). Social Media Analysis using Optimized K-Means Clustering.

[4] Boulder, B., Kissoon-charles, L. T., Based, D., & Toya, L. (2015). On K-Means CLustering Using Mahalanobis Distance, (June).

[5] Cafuta, D., & Dodig, I. (2017). Spam Detection Based on Search Engine Creditworthiness, 135–141.

[6] Crc, H., & Hofmann, M. (2016). Graph-Based Social Media Analysis.

[7] Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[8] Domínguez, D. R., Díaz Redondo, R. P., Vilas, A. F., & Khalifa, M. Ben. (2017). Sensing the city with Instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, *78*, 319–333. https://doi.org/10.1016/j.eswa.2017.02.018

[9] Hong, S., Lee, W., & Han, M. (2015). The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification, *7*(1).

[10] Hu, L., Gao, W., Zhao, K., Zhang, P., & Wang, F. (2018). Feature selection considering two types of feature relevancy and feature interdependency. *Expert Systems with Applications*, *93*, 423–434. https://doi.org/10.1016/j.eswa.2017.10.016

[11] Kaplan, A. M., & Haenlein, M. (2010). The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

[12] Kaur, R., Singh, S., & Kumar, H. (2018). Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches. *Journal of Network and Computer Applications*, *112*, 53–88. https://doi.org/10.1016/j.jnca.2018.03.015

[13] Kawash, J. (n.d.). *Online Social Media Analysis and Visualization*.

[14] Kong, X., Hu, C., & Duan, Z. (2017). *Principal Component Analysis Networks and Algorithms*. https://doi.org/10.1007/978-981-10-2915-8

[15] Mousavi, M., Bakar, A. A., & Vakilian, M. (2015). Data stream clustering algorithms: A review. *International Journal of Advances in Soft Computing and Its Applications*, *7*(Specialissue3), 1–15.

[16] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

[17] Peng, S., Zhou, Y., Cao, L., Yu, S., Niu, J., & Jia, W. (2018). Influence analysis in social networks: A survey. *Journal of Network and Computer Applications*. https://doi.org/10.1016/j.jnca.2018.01.005

[18] Prasetijo, A. B., Isnanto, R. R., Eridani, D., Alvin, Y., Soetrisno, A., Arfan, M., & Sofwan, A. (2017). Hoax Detection System on Indonesian News Sites Based on Text Classification using SVM and SGD, 45–49. https://doi.org/10.1109/ICITACEE.2017.8257673

[19] Rahim, N. A. A., & Sulaiman, S. (2015). Social network analysis for political blogosphere dataset. *International Journal of Advances in Soft Computing and Its Applications*, *7*(Specialissue3).

[20] Sedhai, S., & Sun, A. (2018). Semi-Supervised Spam Detection in Twitter Stream. *IEEE Transactions on Computational Social Systems, 5*(1), 169–175. https://doi.org/10.1109/TCSS.2017.2773581

[21] Shehnepoor, S., Salehi, M., Farahbakhsh, R., & Crespi, N. (2017). NetSpam : a Network-based Spam Detection Framework for Reviews in Online Social Media, *6013*(i), 1–10. https://doi.org/10.1109/TIFS.2017.2675361

[22] Tharwat, A. (2016). Principal component analysis - a tutorial. *International Journal of Applied Pattern Recognition*, *3*(3), 197. https://doi.org/10.1504/IJAPR.2016.079733

[23] Tsvetovat, M., & Kouznetsov, A. (2011). *Social Network Analysis for Startups*. https://doi.org/10.1017/CBO9781107415324.004

[24] Yu, D., Chen, N., Jiang, F., Fu, B., & Qin, A. (2017). Constrained NMF-based Semi-supervised Learning for Social Media Spammer Detection Dingguo. *Knowledge-Based Systems*. https://doi.org/10.1016/j.knosys.2017.03.025