Journal of Computing and Applied Informatics (JoCAI) Vol. 4, No. 2, 2020 | 111 – 121

talenta PUBLISHER

DATA SCIENCE
Journal of Computing and Applied Informatics

Cluster Analysis of Online Shop Product Reviews Using K-Means Clustering

***Rena Nainggolan*[1], *Eviyanti Purba*[2]**

[1,2]*Universitas Methodist Indonesia, Faculty of Komputerisasi Akuntansi, Medan, Indonesia*

Abstract. Technological developments have made changes in people's lifestyles, namely changes in the behavior of people who had shopped directly or offline to online. Many benefits are obtained from shopping online, namely the many conveniences offered by shopping online, besides that there are also many disadvantages of shopping online, namely the many risks in using e-commerce facilities, namely the problem of product or service quality, safety in payments, fraud. This research aims to mine review data on one of the e-commerce sites which ultimately produces clusters using the K-Means Clustering algorithm that can help potential customers to make a decision before deciding to buy a product or service

Keyword: *Data Mining, K-Means Clustering, Cluster, Online Customer Reviews*

Received 19 November 2019 | Revised 13 March 2020 | Accepted 15 May 2020

1 Introduction

Based on a survey in 2016 APJII (Association of Indonesian Internet Service Providers), stated that Indonesian Internet service users reached 132.7 million of Indonesia's population of 256.2 million. Based on APJII data, 63.1 million people use mobile phones for the internet. And 92.8 million people use mobile internet access internet services. And the site most frequently visited is the online shop shopping site, which amounts to 82.2 million people. [1] The development of e-commerce can not be separated from internet technology. Between 2012-2015 Integrated Community Development (ICD) research institutions. The use of e-commerce increased by 42%. This figure is higher compared to other countries such as Malaysia (14%), Thailand (22%), and the Philippines (28%) [2] The development of online shops has increased greatly in Indonesia, even in remote areas. With the many conveniences obtained by consumers in shopping online makes consumers switch to using these facilities. People only need internet subscription fees to get these facilities. In an online store we will find a lot of sellers, making it easier for prospective

*Corresponding author at: Department of Accounting Computational, Faculty of Economy, Universitas Methodist Indonesia, Medan, Indonesia

E-mail address: renanain99olan@gmail.com

ISSN: 2580-6769 | e-ISSN: 2580-829X | DOI: https://doi.org/10.32734/jocai.v4.i2-2855
Journal Homepage: https://jocai.usu.ac.id

customers to choose the products / services offered by the seller according to their needs, this makes competition among the providers of products / services compete fiercely, it becomes an advantage for prospective buyers, but there are some obstacles in shopping online, before deciding to buy a product / service, prospective buyers must pay attention to the history of the seller whether it can be trusted or not, for that in conducting transactions online, trust is needed between the seller and the buyer, one of factors that greatly affect the consumer to buy goods is to know the history of the seller and how the products offered by the seller, this can be known by the seller by looking at reviews of the product that can be read by prospective buyers through existing product reviews on the online store site.

2 Data Mining

Data Mining aims to extract information and knowledge in proving the accuracy and potential Useful for decision making and problem solving. In general, data mining discusses methods such as classification, regression, variable selection an clustering. Based on the task that can be performed data mining is devide into several sections: [3]

- Description provide possible explanations for a pattern. To describe the pattern and trends contained in the data.
- Estimation The model is built using a complete record that provides the values of the target variable as a predictive value.
- Prediction is classification and estimation are almost the same as predictions, predictions of the value of future results.
- Classification in the classification there are target variable categories. For example. Income classification can be separated into several categories, namely low income, medium income and high income.
- Clustering grouping data/objects into other clusters that have different charateristics and ultimately will produce a cluster or group that has a very high level of similarity

3 Clustering

Clustering is an activity to divide a number of data into groups based on similarities that have been determined beforehand. Clusters are groups or groups of data objects that are smilar to each other in the same cluster and similar to different cluster objects.

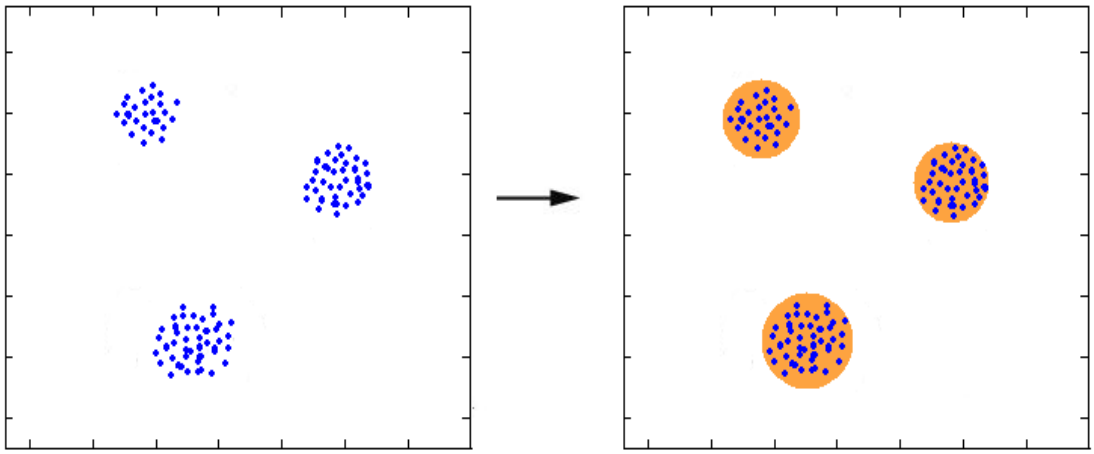

Figure 1 Example of the Clustering Process

3.1 K-Means Clustering

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster. [4]

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters. Method:

- Arbitrarily choose k objects from D as the initial cluster centers
- Repeat
- (re)assign each object to the cluster to which the object is the most similar
- based on the mean value of the objects in the cluster
- Update the cluster means, that is, calculate the mean value of the objects for
- Each cluster
- Until no changes

4 Methodology

4.1 Pre Process Data

In general there are four stages completed in this research model, namely the stage of collecting product review data, the stage of preprocessing data, the feature selection stage, testing the MK-Means clustering model and testing the clustering performance.

A. Data collection

Data collection (data crawling) aims to capture product review data. This research uses product review data obtained from online buying and selling sites. Data is collected using Octoparse application, which is an open source application for web crawlers [5]

B. Text pre-processing

Initial processing or text pre-processing is the second stage in text mining [6]. The initial processing phase aims to prepare data to be used at the pattern discovery stage, for example eliminating data that contains noise, incomplete data and inconsistent data.

C. Feature Selection

At this stage there are two processes that are carried out, as follows:

Case Folding

Case Folding is the process of changing uppercase letters into all lowercase.

Input:

Saya bErmain Lomba Galah

BUK SAYA MAU BERANGKAT

Output :

saya bermain lomba galah

pak mau lapor

Non Alphanumeric Removal

Removes all non-alphanumeric characters. Recall that alphanumerics are letters Sand numbers. Also notice that this may split a token into multiple tokens

Input:

Saya bermain... tadi

TOLONG DIBERSIHKAN KAMAR!!! Kapan siapnya??!?

Output:

saya bermain tadi

TOLONG DIBERSIHKAN KAMAR kapan siapnya

Own Stop Words Removal

You need to provide a list of stop words, then it will be removed from your document. The list of stop words must be placed in a text file, each word in a line. Stop word can also be regular expression, but it must not contain space.

Input:

harga bawang Rp 15.000,00

harga bawang rp 15.000,00

listrik koq naik, warga sedih #edisicurhat

sudah blokir situs http://www.lucu.com

Output

harga cabai Rp 15.000,00

harga cabai 15.000,00

listrik naik, warga sedih

sudah blokir situs

Stop Words Removal

This task includes case folding and remove non-alphanumeric characters. Be warned, "tidak" (not) is also removed. Depending on what you are going to do next, removing this word may affect the resul

Input:

Pak kepala lurah tidak tahu bahwa 4 penari

di rumah itu adalah teman lamanya!

Output

pak kepala lurah tahu 4 penari

rumah teman

Stem

This task includes case folding

Input:

Mempermainkan peranan sebagai putri salju

Output

main peran putri salju

D. Stopword Removal

In the process of stop words, it takes a data or a list of words that you want to be deleted, in general stop words are common words that do not have the meaning namu often appear. In Indonesian such as "to", "with", "which", "if", "will" and so forth. For this reason, a removal is needed.

E. Stemming

Stemming in this research is based on Nazief and Andriani's algorithm. This algorithm is also known as the confix stripping algorithm, which is a special algorithm for stemming Indonesian texts [7].

After all data has been transformed into numerical form, then the data can be grouped using the K-Mean Clustering method. To be able to group these data into clusters, several steps need to be done, namely:

- Determine the desired number of clusters in advance. In this study the existing data will be grouped into two clusters.

- Determine the starting point of each cluster. In this study the initial center point is generated randomly. The center of the cluster in the initial solution can be seen in table 1

5 Result and Discussion

5.1 Results of data collection

Research conducted using online customer reviews consisting of 888 data. From 888 data, there were 806 positive comments and 82 negative comments.

1. bagus bagus bagus bungkus rapi kirim cepat alhamdulillah pas si terimakasih lazada
2. bagus sepatu murah murah kualitas bagus agk longgar anak sm2 ukur cepat 221
3. sepatu bagus kirim cepat terimakasih tukar nomor 35 prosedur
4. kualitas bagus nyaman pakai moga awet
5. bahan bagus besar untung retur ganti ukur
6. sesuai skripsi nyaman moga awet mantap dahh
7. mantap barang cepat barang bagus layan bagus
8. barang terima
9. sesuai gambar
10.bagus bahan sesuai gambar
11.bagus…!
12.terimakasih barang tidak kecewa
13.beda gambar sama barang yang datang…!
14.BARANG JELEK! ☹
15.barang besar puas terimakasih lazada

5.2 Pre Processing Data

After the product review data has been carried out, the next step is so that online customer review data can be applied to the k-means clustering algorithm, then the pre-process data is carried out. The pre-processing stages that are applied are Case Folding, Non Alpha Numeric Removal, Stop words Removal, and Stemming. The list of stop words for Indonesian consists of 760 words [8]. The Stemming Algorithm that is applied is a special stemming algorithm for the Indonesian]language, the nazief-Andriani algorithm [9]

1. Kemas hancur
2. barang lama packing hancur
3. barang datang lama
4. kecewa tidak sesuai harap
5. barang pesan
6. barang tidak sesuai deskripsi
7. barang pesan sekarang
8. pesan
9. tidak sesuai ukuranya
10.lama sampai

A. String to word vector

To convert string data into word vectors, the TF-IDF algorithm is applied. Results of Implementation of TF_IDF produces a data matrix with dimensions of 85 attributes x 888 data. There are 85 terms in the data as shown in the following figure.

Alhamdulilah, aman, aneh, apik, awet, bagus, baik, batal, beda, bekas, besar, berar, besar, bingung, bohong, bolong, bongsor, buruk, cacat, cantik, cepat, cocok, elegan, enak, enteng, ganti, hancur, jangkau, jelas, jelek, kacau, kasar, kece, kecewa, kecil, keras, keren, kesal, kilat, komplit, lama, lambat, lembut, lengkap, lentur, lepas, licin. Longgar, lumayan, mahal, mantap, mengelupas, mudah, murah, nyaman, nyasar, parah, pas, persis, profesional, ramah, rapi, ribet, ringan, rusak, sakit, salah, sama, sampai, senang, sesuai, simpel, sobek, sopan, suka, sukses, super, tahan, tebal, telat, terimakasih, tidak, tinggal, tipu, trendi, tukar

5.3 Converting Data to Numeric

The following table is an example of the final data from the conversion

{26 3.590179,146 1.41682,281 3.109726,328 1.435142}
{13 0.504275,56 3.944257,102 3.944257,225 1.760924,247 1.994149,297 3.356954}
{13 0.504275,18 4.705757,95 4.705757,123 4.705757,143 3.590179,146 1.41682,225
{3 3.356954,19 1.072454,30 1.903327,76 1.39897,77 4.705757,85 1.928085,93 3.744851,139 2.396048,146 1.41682,204 4.225304,208 3.944257,223 3.590179,243 1.823039,276 0.957791,299 4.705757,324 1.218923,328 1.435142,344 1.479877,351 3.944257}
{5 4.705757,10 2.783945,13 0.504275,14 2.595454,19 1.072454,41 2.202859,146 1.41682,173 2.082756,189 4.705757,210 1.823039,211 3.356954,243 1.823039,269 }
{13 0.504275,19 1.072454,30 1.903327,41 2.202859,

5.4 Attribute Selection

The data above is still too large and ineffective, so the attributes must be filtered. By using the Cfs algorithm, as many as 50 attributes.

Alhamdulilah, aman, aneh, apik, awet, bagus, baik, batal, beda, bekas, besar, berar, besar, bingung, bohong, bolong, bongsor, buruk, cacat, cantik, cepat, cocok, elegan, enak, enteng, ganti, hancur, jangkau, jelas, jelek, kacau, kasar, kece, kecewa, kecil, keras, keren, kesal, kilat, komplit, lama, lambat, lembut, lengkap, lentur, lepas, licin. Longgar, lumayan, mahal, mantap

5.5 Codifier K-Mean Clustering

Table 1 Some examples of Calculation of Mean and Standard Deviation values

Atribut	Nilai	Cluster	
		0	1
alhamdulilah	Mean	0.0483	0
	Std. dev	0.3803	0.3368
Aman	Mean	0.0068	0
	Std. dev	0.1786	0.1579
Aneh	Mean	0.0122	0
	Std. dev	0.2266	0
Apik	Mean	0	0.0242
	Std. dev	0.2579	0.3365
Awet	Mean	0.0041	0.2144
	Std. dev	0.1067	0.7422
Bagus	Mean	0.294	0.0642
	Std. dev	0.2468	0.1681
Baik	Mean	0	0.04334
	Std. dev	0.2004	0.4262
Batal	Mean	0	0.077
	Std. dev	0.2509	0.5315
Beda	Mean	0	0.1208
	Std. dev	0.297	0.6252
Bekas	Mean	0	0.0242
	Std. dev	0.1579	0.3365
Berat	Mean	0	0.4334
	Std. dev	0.2004	0.4264
Besar	Mean	0.1481	0.0296
	Std. dev	0.5099	0.2357
bingung	Mean	0	0.04334
	Std. dev	0.2004	0.4262
Bohong	Mean	0	0.0242
	Std. dev	0.1579	0.3365
……			
Mantap	Mean	0.0337	0.3058
	Std. dev	0.2629	0.7795

5.6 Cluster Instances

Table 2 Cluster Instance

Cluster		Persentase	
0	1	0	1
729	159	82%	18%

5.7 Clustering Process

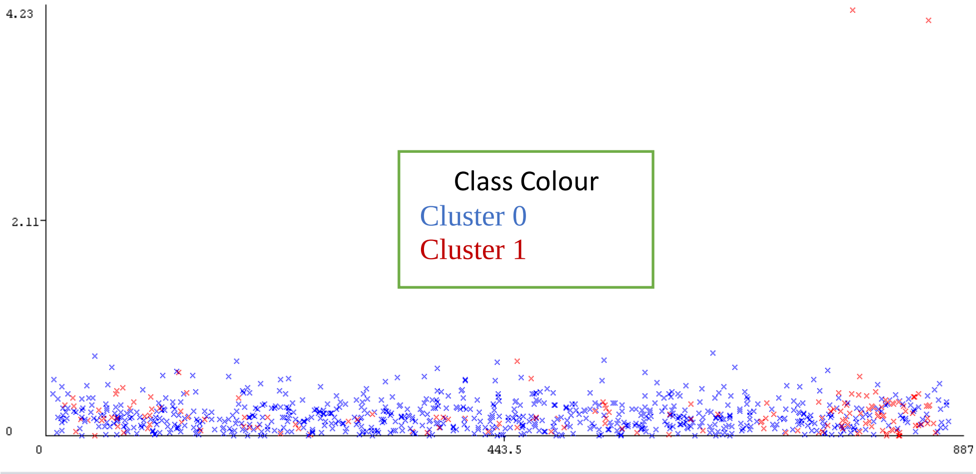

Figure 2 Cluster data

5.8 Some examples of data clustering result

```
0,0,0,0,0,0,0.504275,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.579067,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
2,0,0,0,0,0,0.504275,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.579067,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
4,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.845045,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3.463804,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.845045,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.966963,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster1
7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
8,0,0,0,0,0,0.504275,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
9,0,0,0,0,0,0.504275,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.845045,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
10,0,0,0,0,0,0.504275,0,0,0,0,0,1.903327,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
11,0,0,0,0,0,0.504275,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.579067,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
12,0,0,0,0,0,0.504275,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.579067,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
13,0,0,0,0,0,0.504275,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
14,0,0,0,0,0,0.504275,0,0,0,0,0,1.903327,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,cluster0
```

5.9 Graph Comparison of cluster 1 and cluster 2 data can be shown in the following graph

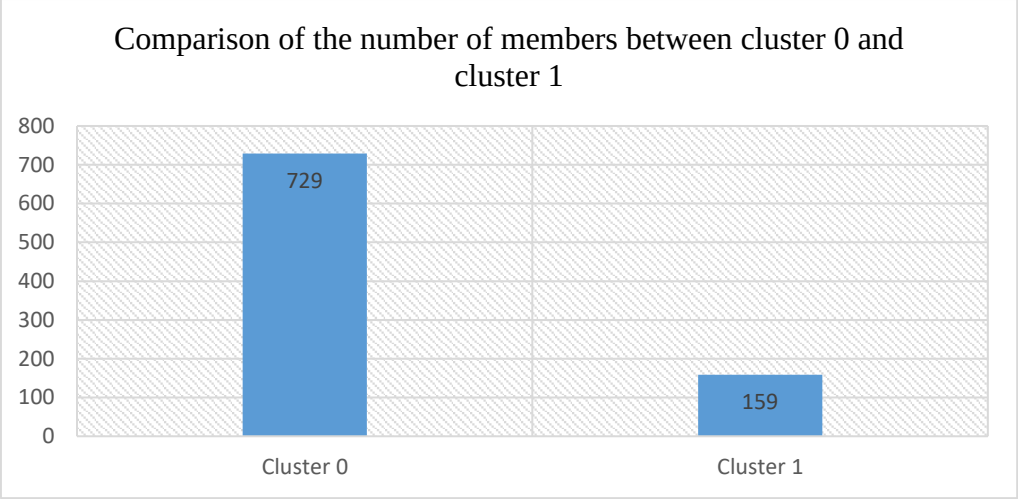

Figure 3 Graph Comparison of the number of members between cluster 0 and cluster 1

6 Conclusion

From table.2 above it can be concluded that the results of testing of 888 reviews produced 2 clusters, namely:

1. Cluster 1, which produces 729 (82%) reviews that have a very high similarity grouped into 1 cluster
2. Cluster 2 produced 159 (18%) reviews that had very high similarities grouped into 1 cluster group.

REFERENCES

[1] I. U. Yoviriska, and Wahjoedi. "Trend keputusan Belanja Online Mahasiswa Fakultas Ekonomi UM Angkatan 2014", *Jurnal Pendidikan Ekonomi*, vol.11, no. 1, Mar. 2018.

[2] S. Sidharta, and B. Suzanto, "Pengaruh kepuasan transaksi online shopping dan kepercayaan konsumen terhadap sikap serta perilaku konsumen pada e-commerce", *Jurnal computech & Bisnis*, vol. 9, pp. 23-26, no.1, Jun. 2018.

[3] V. Carlo. *Business Intelligence: Data Mining and Optimazation for Decision Making*, West Sussex, United Kingdom: John Wiley & Sons Ltd, 2009.

[4] R. Nainggolan and E. Purba, "The Cluster Analysis of Online Shop Product Reviews Using K-Means Clustering", *Data Science: J. of Computing and Appl. Informatics*, vol. 4, no. 2, Jul. 2020.

[5] Y. Ganjisaffar. (2013). *Open Source Web Crawler for Java*. [Online]. Accessed: 13 May 2017. Available: http://code.google.com/p/crawler4j/.

[6] L. Kumar, and P. K. Bhatia, "Text Mining: Concepts, Process and Applications", *Journal of Global Research in Computer Science*, vol. 4, pp. 36-39, 2013.

[7] T. Mardiana , T. B. Adji, , and I. Hidayah, "Stemming Influence on Similarity Detection of Abstract Written in Indonesia", *TELKOMNIKA* (*Telecommunication Computing Electronics and Control*), vol. 14, no. 1, pp. 219-227, 2016.

[8] F.Z. Tala, "A Study of stemming effects on information retrieval in Bahasa Indonesia", Master of Logic Project, Institute for Logic, Language and Computation, Univ. van Amsterdam, Netherlands, 2013.

[9] M. Adriani, J. Asian, B. A. A. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A confix-stripping approach", *ACM Transactions on Asian Language Information Processing*, vol. 6, no. 4, pp. 1-33, 2007.