Comparison of Different Types of Cluster Algorithm on Netherland Telecommunication Provider Solutions

Christophorus Jeremy Wicaksana
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands
christophorusjeremywicaksana@student.utwente.nl

ABSTRACT

Clustering is an unsupervised method of grouping data into different groups. In this research the data is in a form of text. The text then needs to be translated into numbers with word2vec skip-gram. The numerical data created will be used in the 4 different types of clustering algorithm which is divided into different groups based on the numerical data based on sample data of Netherland's telecommunication regarding the solution in telecommunication problem. This research is conducted because there has not been any research on comparing clustering algorithm in telecommunication data and there is no research on comparing 4 types of clustering algorithm on text data. So, 4 types of clustering algorithm which are Centroid-based clustering, Density-based Clustering, Hierarchical Clustering, and Distribution-based Clustering used for clustering the text documents. The result then is evaluated by S_Dbw and observation of the result from each algorithm. The result of the research showed that centroid-based has the best performance compared to other clustering algorithms based on the telecommunication data while Hierarchical algorithm is the most consistent on grouping the data.

Keywords

Preprocessing, k-medoids, OPTICS, Hierarchical, GMM, S_Dbw

1. INTRODUCTION

Telecommunication infrastructure and internet are part of daily needs for everyone and the foundation to keep society running properly. So, telecommunication provider needs to be able to maintain major problem such as outage or maintain availability of contact emergencies in Netherlands, 112. If the provider manages to solve the major problem, then they need to submit their solution to the national providers [6]. However, the solution needs to be anonymized such that it does not contain any data related to the origin of the providers due to commercial reason. Because of that the research is going to use the sample in the form of anonymized data. The data used will be implemented on unsupervised machine learning given the fact no labels on the data in order to help the telecommunication providers categorized the data which in this case going to use clustering without doing it manually.

35th Twente Student Conference on IT, April. 26th, 2021, Enschede, The Netherlands.

In this research there are limitation on the cleaning the data because some words still in non-basic form. For example, "monitoring" is not lemmatized (reduce the word into its basic form) into "monitor" while there is other sentence that contain word "monitor". Since the data is high dimension because of its features based on every unique word the only way to reduce the data is by using PCA where it is not 100% accurate and the plot diagram is hardly observable. The data is in the form of text, it needs to be translated into numbers in order the machine learning can understand. The translation algorithm from text to number which in this case is word2vec is not able to put the similar word into same weight because every word is deemed as different words so word "like" and "liking" will be different. Lastly, this research does not have any ground truth on the clustering, so it depends on cluster validation S_Dbw score which see how dense and scatter the clusters are and a observation on the data.

There are many clustering algorithms exist, but there has not been a comparison on clustering algorithm based on its type while there are many other paper comparing between 2 clustering algorithm without any basis on why it is chosen [16,7,13]. Also there has not been any research on telecommunication provider solution on major problems by using clustering techniques. The research is mainly focusing on what is the best out of 4 types of clustering algorithm based on google developers[1] which are Centroid-based clustering, Density-based Clustering, Hierarchical Clustering, and Distribution-based Clustering. The result will be based on the S_Dbw score and the observation on the data.

There are 2 main research questions in this research which follow CRISP-DM cycle[2]. First, "what is the appropriate number of clusters for each algorithm?" this question objective is trying to find the best cluster performance from each clustering algorithm. The comparison is within the same range to each other to ensure the validity of the result. In order to find the best cluster, the S_Dbw score is used as the comparison. The second research question is "which is the best clustering algorithm based on the possible factors as comparison" and the reason as of why the clustering algorithm is the best.

The paper is structured into 4 sections with abstract and references. The first section is the introduction explaining what the background is, limitation, problems, research questions, and the structure of the paper also there is research question in subsection 1. In the section 2 is the background of the research all of the information that supports the research and why it is chosen. Section 3 is the findings of the report that shows what is

[1] https://developers.google.com/machine-learning/clustering/clustering-algorithm

[2] https://www.datascience-pm.com/crisp-dm-2/

the result and what has been done in the whole of the research. Lastly, section 4 is the conclusion to conclude the paper.

1.1 Research Questions

Based on CRISP-DM process model. There are 2 main research question to be answered at the end of this paper. The question is focusing on data modelling and data evaluation. The other CRISP-DM cycle will be explained in methodologies. The provider solutions refer to the data that will be used in this research.

Main research question 1:

What is the appropriate number of cluster groups for each algorithm?

- What are the appropriate range value of clusters for each algorithm except density-based ratio?
- What is the lowest cluster validation value for each algorithm?
- Will every/any algorithm have the same amount of cluster groups based on lowest cluster validation value of S_Dbw?

Main research question 2:

Based on 4 types of clustering algorithm which one is the most optimum implementation for provider solutions?

- What factors from each algorithm can be used as comparison?
- What causes each algorithm is better than others?

2. STATE OF ART

2.1 Preprocessing Text

The subsection is discussing steps on how to clean the data. Preprocessing text data is a method to clean the data and always used in text mining [1]. This is needed in order to maintain consistency of text data that is translated into statistical feature in form of numbers.

There are 5 steps needed to be performed on the raw data before applied to the algorithm [9,19,2]. Lower casing the words to prevent any capitals in the text. Tokenizing the texts into words so it can be processed in the next step. Removing punctuation and then stopwords because clustering is sensitive to stopwords and lastly implementing lemmatization which generalize the words in the same basic form. For instance, making plural words into base form which is singular.

2.2 Word2vec

Word2vec is a method to translate the text documents into number that can be used for the clustering algorithm as the inputs.

There are Continuous bag of words (CBOW) and Continuous skip gram model in word2vec [12]. Here the Continuous skip gram model is implemented to weight text documentation data into numbers that can be used in clustering algorithms. This algorithm is chosen because based on Thomas Mikolov research itself skip gram model is better with higher accuracy algorithm compared to CBOW. Based on the picture above it is the architecture of the Continuous skip gram model which is an algorithm to predict a word based on neighbor words in the sentence [12]. In this case, Mikolov use negative sampling model which could differentiate the noises and the actual data to be processed. In the Figure 1 it can be seen clearly the differencs between CBOW and Skip-gram.

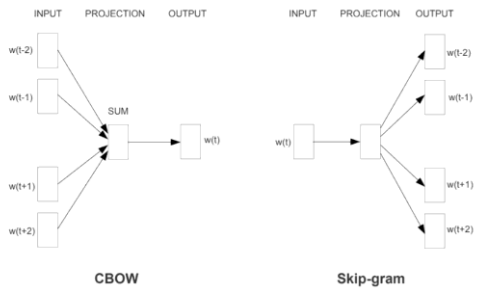


Figure 1. Workflow of CBOW and Skip-gram

2.3 Clustering

Clustering is an unsupervised learning method that can classify data into different clusters of data. Not only numerical data, but text documents also possible to be clustered. However, there is no way for clustering to understand documents directly, the documents need to be translated into statistical features through a weighted text classification in a form of number. In this research, Word2vec is used for that purpose which is previously explained in detail in section 2.2.

There are many clustering methods available, but in this research 4 algorithms are chosen because based on google developers[1] there are 4 types of clustering. Centroid based clustering, in this case k-medoids is used. Density-based clustering which uses the OPTICS algorithm. Distributed-based clustering, Gaussian Mixture Model. Lastly, Hierarchical clustering using agglomerative hierarchical. Those algorithms used to represent each type will be explained later in detail.

2.3.1 K-medoids

As the name implies K-medoids is the combination of K-means concept and using medoids instead of means for reference points of a cluster [18]. Medoid is a statistical function to find a central point of the data which can be seen on equation below. Later, the reference point from the equation helps reducing sensitivity of an outlier in set of data. For instance, an outlier will be nearer from the reference point rather than using means, so it minimizes dissimilarities between data [4]. The algorithm of K-medoids is similar to K-means it chooses random data as medoid points. Other points then see the nearest reference points to each other. After that all data now in clusters. Now each cluster try to find medoids for each cluster then iterate the whole data points based on the medoids value hoping for changes. If there are no changes then the clustering is final. Based on how many iterations it wants, the algorithm will store the clustering results as many as given iteration then see which result is the most balanced clustering.

$$x_{\text{medoid}} = \arg\min_{y \in \mathcal{X}} \sum_{i=1}^{n} d(y, x_i)$$

2.3.2 Agglomerative Hierarchical

Hierarchical algorithms produce a tree-like hierarchy graph to search the data within. In this case the agglomerative method is implemented. Agglomerative Hierarchical is using a bottom-up approach where it standardizes every data into individual clusters and then merges every similar group into the same cluster [5] by using distance metrics similarly to k-medoids implementation where in this case Euclidean is used again. There are 4 merging methods available from agglomerative algorithm: Single linkage clustering, Group-average linkage clustering, complete linkage clustering and lastly ward's method. In this research, ward's method is chosen because its

ability to allow cluster center to be specified and using error sum square equation to see how similar each data is [14]

2.3.3 OPTICS

Optics algorithm is an extended DBSCAN algorithm where it can produce not only clusters but also noises if current data point is located out of the given parameters of minimum number of object within the minimum points(minpts) and radius ε. If a chosen neighborhood-ε manages to find another neighborhood-ε within the radius and at least the data in ε is at least minpts then a new cluster with ε-neighborhood is created [5]. And the current neighborhood tries to find another neighborhood until there are no more. However, the method is mostly based on DBSCAN. What differs from OPTICS is the ability of OPTICS algorithm to automatically find the best distance ε based on the minipts given as long as the ε is smaller than the borderline of the data. [3]

2.3.4 Gaussian Mixture Model

As the name suggest it uses Gaussian distribution or usually called Normal distribution. The idea is to find all the probability of the datasets and put it into the proper cluster group. Since it is probability, the data can be in more than one cluster. However, in this research hard clustering is used so it will choose the highest probability cluster groups for the data and set it to the groups. [15]

Figure

$$N(X|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}\sqrt{|\Sigma|}} exp\left\{-\frac{(X-\mu)^T\Sigma^{-1}(X-\mu)}{2}\right\}$$

In the figure(above) X represent the current data being observed, μ is the mean, and Σ is the covariance matrix in which can described 4 shapes of matrix: full, diag, tied, and spherical based on scikit GMM[3]. Based on scikit tied shapes is the best shape compared to others with accuracy of 95%.[4]

2.4 Internal Cluster Validation

Evaluation of the documents grouped by clustering algorithms is based on how dissimilar the content of each cluster is compared to others. This can be achieved manually to see how unique each cluster is, but it is exhausting if the result to be evaluated is extremely large. Fortunately, there is an autonomous method to validate the clustering algorithm based on the distance of each cluster. There are 2 types of cluster validation, external (based on predefined structure) and internal validation (based on the current data) where both of them are considered compactness (how close each cluster is) and separability (how distinct between each cluster) of the data.

Based on Rejito [17] internal cluster validation method is far more superior than external cluster validation where internal clustering validity reached 86% accuracy compared to external method with 51.9% accuracy based on k-means algorithm.

In internal validation method there are 11 different measures [11] and S_Dbw validity index is the best measure out of all 11 based on Liu. It is because of its consistent performance in different aspects of data. Monotonicity, which is well separated data, noise if there is data out of range from any clusters, density where the data is mostly scattered around without the noises, and lastly subclusters in which the clusters are closed to each other are the aspects experimented. S_Dbw manages to always get the best results in every aspect.

2.4.1 S_Dbw validity index

As mentioned in section 2.4 validity index is considering separation and compactness of the clusters for the validation criteria. Compactness of the cluster is defined as the density relation among the clusters. Compactness of the cluster evaluates the average density of among of the clusters by using below equation

Figure

$$Dens\_bw(c) = \frac{1}{c\cdot(c-1)}\sum_{i=1}^{c}\left(\sum_{\substack{j=1\\ i\neq j}}^{c}\frac{density(u_{ij})}{\max\{density(v_i), density(v_j)\}}\right)$$

Where u_{ij} is the middle point of the line segment created by v_i and v_j cluster centers from cluster c. The density function is defined from accumulation from f(x,u) by the total number of tuples n_{ij} times.[8,10]

$$f(x,u) = \begin{cases} 0, & \text{if } d(x,u) > stdev \\ 1, & \text{otherwise} \end{cases}$$

For separation, it finds the average of the scattered data in the clusters. It is defined purely by mathematical equations by using variance. The equation is defined below

$$Scat(c) = \frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\| \Big/ \|\sigma(S)\|$$

Lastly the S_Dbw(c) for cluster c result is defined by addition compactness and separation of the cluster where the smaller the result is the better the clustering is because it shows that the cluster is neither compact nor scattered in the graph. The equation can be shown below

$$S\_Dbw(c) = Scat(c) + Dens\_bw(c)$$

3. METHODOLOGIES

The number of data used in this experiment is 261 data consisting of duplicates data, on the other hand there are 179 data if the duplicates are removed. Both datasets are going through same process to be used by the algorithm in the end. First the data is cleaned through preprocessing text where it removes stopwords, punctuation, generalize the words, and lower case all of the words that is already tokenized. The cleaned text data then will be translated into matrices of number based on the unique words from the whole documents where the matrices of data generated through word2vec which consist of 484 unique feature for each sentence in the documents. The matrix is then being used as an input for each of the 4 different types of clustering algorithm where all of the algorithms will have its own S_Dbw score. The OPTICS algorithm is experimented first to get the best cluster group based on the S_Dbw because every algorithm except OPTICS has a expected cluster group as the parameter. So, OPTICS will get the best cluster group and the rest of the algorithm used the cluster group where it is within the range of OPTICS best cluster groups which in this case the interval is +5 and -5 from OPTICS cluster groups. The algorithms now have different S_Dbw results, and it will be compared based on which one has the lowest score. Additionally, there will be an observation by creating a ground truth based on the sentence with topic about "crisis organization". Each algorithm then will be compared which one is the best on clustering the group on "crisis organization" topics.

[3] https://scikit-learn.org/stable/modules/mixture.html#mixture

4. FINDINGS

4.1 Choosing the algorithm

4.1.1 Choosing Cluster validation

Based on section 2.4 it explains in detail on why S_Dbw is chosen instead of others cluster validation.

4.1.2 Choosing GMM Shape

Table 1. Show the differences of GMM shape based on Skip-gram on duplicates documents.

Clusters and S_Dbw value		
GMM Covariance shape		Word2vec Skip-gram
		Duplicates
Tied	Cluster	**44**
	S_Dbw	**0.320923**
Spherical	Cluster	46
	S_Dbw	0.412933
Full	Cluster	44
	S_Dbw	0.323084
diag	Cluster	45
	S_Dbw	0.408739

There are 4 different matrix shapes available on GMM: tied, spherical, full, and lastly diag. There is no explanation which one is better because of it is only a shape. So, the best shape is based on how the cluster data distribution is. In the figure X above different GMM algorithm with different covariance shape is conducted. Based on the result Tied shape is the best shape with the lowest S_Dbw score. Tied shape means that the cluster boundary shape created between clusters are the same.

4.1.3 Choosing Text Weighting Algorithm

All the table below have different Text weighting algorithm (TFIDF, CBOW, Skip-gram) and compared each Text weighting algorithm with duplicates and without duplicates (no duplicates) documents. Duplicates documents means some sentence is duplicated. On the other hand, without duplicates documents is the vice versa. The algorithms chosen is also based on section 2.3. Here TFIDF (Term frequency Inverse Document Frequency) is a basic BOW (bag-of-words) where it just calculates how many numbers are there in each document and weight the word based on how relevance it is.

Table 2 Duplicates and Without duplicates on TFIDF

Clusters and S_Dbw value			
Algorithms		TFIDF	
		Duplicates	Without Duplicates
K-medoids	Cluster	69	**45**
	S_Dbw	0.467013	**0.640734**
Agglomerative hierarchical	Cluster	**63**	45
	S_Dbw	**0.394624**	0.66537
GMM(Tied)	Cluster	69	45
	S_Dbw	0.461706	0.684143
OPTICS	Cluster	65	41
	S_Dbw	0.493382	0.724897

Table 3. Duplicates and Without duplicates on CBOW

Clusters and S_Dbw value			
Algorithms		Word2vec CBOW	
		Duplicates	Without Duplicates
K-medoids	Cluster	**42**	20
	S_Dbw	**0.32526**	0.490663
Agglomerative hierarchical	Cluster	37	16
	S_Dbw	0.353083	0.459175
GMM(Tied)	Cluster	43	21
	S_Dbw	0.361534	0.746275
OPTICS	Cluster	41	**25**
	S_Dbw	0.414534	**0.453449**

Table 4. Duplicates and Without duplicates on Skip-gram

Clusters and S_Dbw value			
Algorithms		Word2vec Skip-gram	
		Duplicates	Without Duplicates
K-medoids	Cluster	44	**25**
	S_Dbw	0.329477	**0.405646**
Agglomerative hierarchical	Cluster	40	16
	S_Dbw	0.334669	0.456749
GMM(Tied)	Cluster	**44**	17
	S_Dbw	**0.320923**	0.595597
OPTICS	Cluster	42	21
	S_Dbw	0.343295	0.746613

The table 2, 3, and 4 is using same clustering algorithm, but use different text weighting algorithm: TFIDF, CBOW, and Skip-gram. The algorithm actually can be generalized into just word frequency (TFIDF) and word embedding (CBOW and Skip-gram) where it looks the relation of the each words within the sentence.

Due to the sample data is given with some duplicates, the experiment is divided into duplicates and without duplicates documents and evaluate the cluster based on S_Dbw. The lowest S_Dbw value means the data in the cluster is more compact. The result shows that in all of cases the duplicates documents are better than without duplicates because of the weight of the words in duplicates is more various due to multiple same documents which increase the score of that duplicates documents. On the other hand, without duplicates documents are more scattered because there are some words that are in same but in different sentences.

In almost all the cases TFIDF always perform the worst based on the S_Dbw value. This is because TFIDF only looks the word itself and see how the word is relevant compared to other documents. Addition observation to the argument why TFIDF is not recommended because using TFIDF resulting the clustering of the data is too loose and not too restricted. That means the sentences will be in a cluster with other sentences that does not have any relation at all. For example, in the figure x below the sentences in line 4 and line 5 has the word “give” in the

beginning but does not necessarily mean the sentence has the same meaning. As it can be seen, line 4 talks about monitor crisis improvement while in line 5 it talks about influence influencers. This happens because TFIDF only see the similar words in the sentences. Each sentence has the word "give" in the beginning and the rest of the words is different.

	cluster	data
2	0	create function description al crt chair information manager advisor channel coordinator factcheckers
3	0	give early information cause disturbance keep communicating
4	0	give one central part mandate overview monitor crisis improvement initiative
5	0	give something influence influencers
6	0	good relationship press
7	0	influence press giving extra information future...
8	0	make information manager responsible maintaining date contactlists scheduals
9	0	make sure cmt core resoution team get access required tool via information manager
10	1	make influencors part certain crisis communication strategy
11	1	negotiate claim
12	1	prepare certain standard communication text recording advance
13	1	prepare claim mitigation strategy
14	1	prepare standard easy deploy compensation package
15	2	actively communicate status expected resolution time available

Figure 2. Preview of TFIDF clustering on the documents

In the case of word embeddings as shown in table 3 and 4 it shows that duplicates documents are better in comparison to without duplicates again based on the S_Dbw score. In this case Skip-gram is chosen instead of CBOW because of the S_Dbw in Skip-gram is lower than in CBOW. Besides that, in Skip-gram the S_Dbw never reaches 0.4 in duplicates documents while in CBOW with OPTICS algorithm. Which means Skip-gram is more consistent in comparison to CBOW. In the Figure 3 line 94 and Figure 4 line 3 show the same sentences. However, they are in a different cluster where in CBOW it emphasizes the word "crisis" more and in Skip-gram the word "communication" is emphasized more. Because the result both of the algorithm is different and there is no ground truth to the data. S_Dbw is the only right metrics in this case. So, Skip-gram is preferred.

92	9	establish crisis organisation right mandate procedure resource skill
93	9	establish crisis organisation right mandate procedure resource skill
94	9	involve influencors actor crisis communication
95	9	make crisis organisation member aware crisis tool available phone sims apps ncv
96	9	make crisis organisation member aware crisis tool available phone sims apps ncv
97	9	make crisis organisation member aware crisis tool available phone sims apps
98	9	make crisis organisation member aware crisis tool available phone sims apps
99	9	make crisis organisation member aware crisis tool available phone sims apps
100	9	make crisis organisation member aware crisis tool available phone sims apps
101	9	make crisis organisation member aware crisis tool available phone sims apps
102	9	make crisis organisation member aware crisis tool available phone sims apps
103	9	make hospital aware ncv phone number contact gamma
104	9	make network administration available crisis organisation
105	9	make network administration available crisis organisation
106	9	train staff crisis organisation
107	9	use available tool best effort base
108	10	make awareness campaign discus exemples training session

Figure 3. Implementation with CBOW

	cluster	data
2	0	shadowing exact copy
3	1	involve influencors actor crisis communication
4	1	prepare b2b incident crisis communication process procedure sunny day rainy day
5	1	prepare certain standard communication text recording advance
6	1	strong incident management process
7	1	train arrange fall back
8	1	use latent need part marketing crisis communication
9	2	improve communication response time
10	2	improve communication response time
11	2	improve communication response time
12	2	improve communication response time
13	2	improve communication response time

Figure 4. Implementation with Skip-gram

4.2 Total Cluster Group [RQ1]

Every cluster algorithm except OPTICS is using number of clusters as the parameter of the algorithm because OPTICS algorithm does not use cluster as parameters but minimum steepness on the reachability plot and automatically find the best cluster groups. Because of that, the OPTICS algorithm needs to be conducted before other algorithms and choose the cluster groups based on the lowest S_Dbw score value. The cluster groups from OPTICS then become the reference cluster group as mentioned in section 3.

Based on table 4 in the previous page, OPTICS best cluster group value is 42 with S_Dbw of 0.343295 under duplicate documents. The high value of S_Dbw in OPTICS compare to each other is because in OPTICS there is 1 noise group consisting non-duplicates data and each of duplicates are clustered in the same group so in addition there are 41 more cluster groups..

Given OPTICS best cluster group, the other algorithm needs to find appropriate cluster groups range to be compared in the end where the comparison is sensible enough. Because in the experiment each clustering algorithm beside OPTICS are tested from cluster group value of 2 until 70 and found that every algorithm has lower S_Dbw value at cluster group of 70. The S_Dbw trend to the decrease as the cluster groups increases this is because the data is gradually separated into their own unique cluster groups which loses the point of clustering different data into one cluster groups. However, it is found that starting at cluster groups of 49 the trend is began to show. So, cluster group below 49 is chosen where in the end other algorithm except OPTICS uses cluster group x of $37 < x < 47$ as the appropriate range.

Based on table 4, it can be found that k-medoids have cluster validation value of 0.329477 as the lowest, Hierarchical clustering is 0.334669, GMM is 0.320923, and lastly OPTICS is 0.343295. this shows that OPTICS has the worst cluster validation where in plot diagram on figure 8 shows that the green color is scattered all over the places so it is not a good clustering. While the other plot diagram from figure 5,6, and 7 they have better clustering with less scattered data such as OPTICS that's why OPTICS is within the S_Dbw score of 0.34 while others are below.

Again, in table 4 shows the result only K-medoids and GMM algorithm that has the same cluster of 44. It might be coincidence, however in table 2 it can be seen both are the same. Also, in table 3 the differences between both are 1 cluster. Those similarities shows that K-medoids and GMM has a similar logic where GMM use same logic as K-medoids to generate a cluster points randomly and then only differs on how to compute which data belong to which clusters. On the other hand, OPTICS and Hierarchical agglomerative totally use different algorithm where OPTICS each point is trying to find other points within the given radius while Hierarchical trying to see the similarities between data and merge it into one cluster.

Cluster validation is the only way to choose cluster groups besides manually check every possibility on each algorithm. And based on the previous paragraph in this section it can be concluded that K-medoids best cluster group is 44, Hierarchical is 40, GMM is 44, and lastly OPTICS is 42.

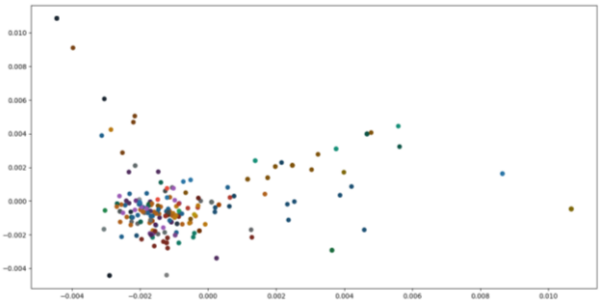


Figure 5. GMM plot under table 3 condition (44 clusters)

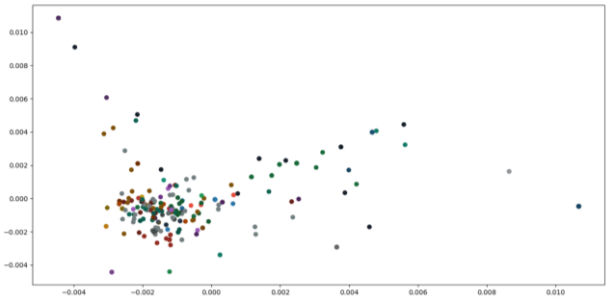


Figure 6. K-medoid plot under table 3 condition (44 clusters)

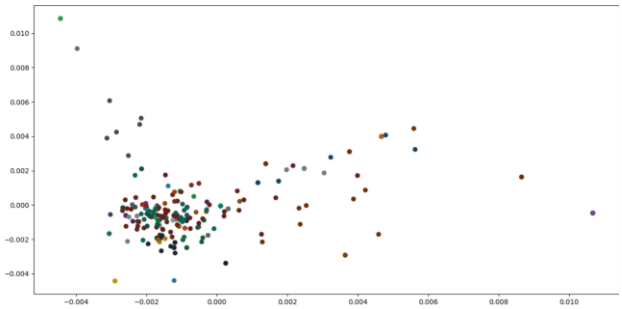


Figure 7. Hierarchical Agglomerative plot under table 3 condition (40 clusters)

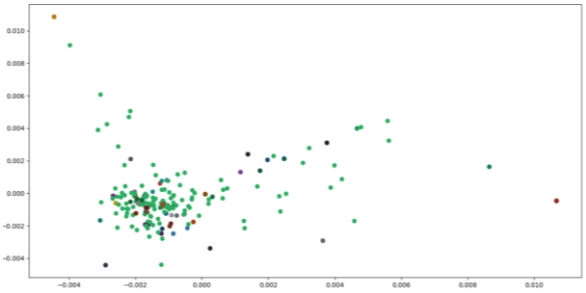


Figure 8. OPTICS plot under table 3 condition (42 clusters) Green color indicates the noise data.

4.3 Best Clustering Algorithm [RQ2]

After thoroughly understand each algorithm there are yet other factor in the algorithm can be compared besides the amount of cluster in each algorithm and manually see the documents clustered in Figure 2, 3, and 4. This is because each algorithm is method is different to each other. For OPTICS algorithm it passes the parameter steepness of the reachability plot where other algorithm can not apply. As for Hierarchical agglomerative has different linkage, in GMM it has different shapes, lastly for K-medoids it passes only cluster group number where it applies for GMM and Hierarchical agglomerative. Hence, there are no other factor can be included in as the comparison for the different cluster algorithm.

Based on table 4 results it can be concluded directly that GMM is better among other algorithm blindly based on the S_Dbw result. However, that reason itself is not enough to prove GMM is the right algorithm. Based on Figure 5, 6 ,7 and 8 the plot diagram is based on 484-dimensional data that is reduced to 2-dimensional data by PCA[5]. So, the representation data points location is not 100% accurate which can be seen in figure 5, 6, 7, and 8.

In OPTICS algorithm figure 8 the green color points that covers everything is the noise data which not included in the clustering data. That means most of the data is being ignored. Based on the observation of the documents, the data is overfitted because the other cluster besides the noise points is only filled by duplicates and each duplicate are on different clusters. Based on the observation 125 documents out of 261 documents are included in the noise cluster group which already rendered half of the documents as useless. So, OPTICS is not a proper algorithm in this case.

As for the rest of the algorithm it is hard to assume which is better just based on the plot diagram. So, in this research a custom ground truth specifically for topic “crisis organization” is created. In the observation the total of whole documents with topic of “crisis organization” is 23 based on Appendix A. Below in figure 9 there are 4 different clustering algorithms. As the previous paragraph said the OPTICS takes some documents into noises and it can be seen in figure 9 because all the data is grouped based on their duplicates and the unique data is grouped as a noise.

Here more observation is conducted for figure 9. Only GMM algorithm divide the “crisis organization” topic into 5 clusters because there is sentence in group 4 based on appendix A figure (a) that has a similar sentence to the group 27. However, the word compared to each other is different and that makes the probability lower, and it is not included together as one group.

On the other hand, K-medoids and hierarchical clustered the documents into 4 clusters. It can be seen directly that K-medoids and hierarchical agglomerative is better because less cluster is created. In figure 9, K-medoids algorithm have the highest cluster group compared to hierarchical which means that k-medoids manage to unify the topic “crisis organization” better than Hierarchical. In the end based on the observation, k-medoids is the best algorithm for this case.

However, the sentence “involve crisis organization” in GMM Appendix A of figure (a) and K-medoids in figure (b) are clustered in the same group, but K-medoids has a larger volume of groups. On the other hand, in Hierarchical figure (d) the sentence is in the whole different cluster groups. In hierarchical it checks all of the available sentences with the current sentence to find the closest numerical value to each other so it is consistent on any cases, but in K-medoids and GMM it generate x random cluster reference points based on how many cluster wants to be created where this is not ideal for any cases because it depends on how good the random cluster point is created which causes the short sentence that did not match with longer sentences grouped together.

So, it can be said that k-medoids is the best clustering algorithm for the data of solution by telecommunication providers in Netherlands. However, Hierarchical is the most consistent algorithm to be used for other different of sample data.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

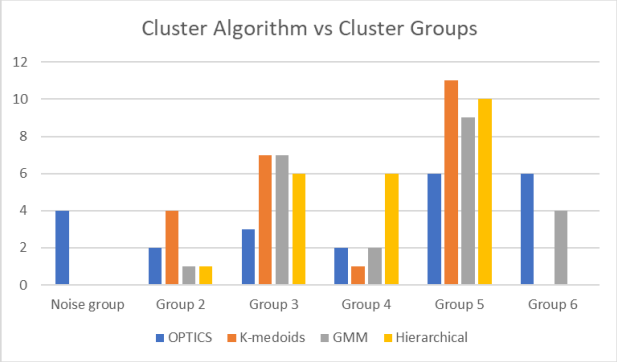


Figure 8. Cluster Algorithm vs Cluster groups visualization based on Appendix A

5. CONCLUSION

This research shown the comparison of 4 types of existing clustering algorithm types based on google developers that density-based clustering is not the best solution because it regards some of the text da ta into noises which means the data is rendered useless and not included in any groups. This is happened because most of the sample data is duplicated so optics try to find the nearest value between the statistical feature of the text data, so it grouped every duplicate into each unique cluster while the other single data is regarded as noises.

On the other hand, K-medoids is the best cluster based on the telecommunication data which is the objective of this research paper while Hierarchical is the best clustering algorithm on similar data that is grouped based on its length which is more consistent rather than k-medoids.

For the future work of this research, if the next researcher wants to conduct this type of research, the research can be improved by using different sets of sample data that contains a ground truth so it can be seen clearly which type of clustering algorithm is the most consistent and the best out of all to use and use sentence2vec to reduce the dimension based on the sentences or using better weighting classifier words algorithm to show a better result when translating the text to statistical number.

6. REFERENCES

[1] Aggarwal, C. C., and Zhai, C. *Mining text data*. Springer Science & Business Media, 2012.

[2] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919* (2017).

[3] Ankerst, M., Breunig, M. M., Kriegel, H.- P., and Sander, J. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record 28*, 2 (1999), 49–60.

[4] Arora, P., Varshney, S., et al. Analysis of kmeans and k medoids algorithm for big data. *Procedia Computer Science 78* (2016), 507–512.

[5] Berkhin, P. A survey of clustering data mining techniques. In *Grouping multidimensional data*. Springer, 2006, pp. 25–71.

[6] Bukhsh, F., Vriezekolk, E., Wienen, H., and Wieringa, R. *Availability Incidents in the Telecommunication Domain: A Literature Review*. DSI technical report series. 2020.

[7] Fauzi, M. Z., Abdullah, A., et al. Clustering of public opinion on natural disasters in Indonesia using dbscan and k-medoids algorithms. In *Journal of Physics: Conference Series* (2021), vol. 1783, IOP Publishing, p. 012016.

[8] Halkidi, M., and Vazirgiannis, M. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings 2001 IEEE international conference on data mining* (2001), IEEE, pp. 187–194.

[9] Hotho, A., Nürnberger, A., and Paaß, G. A brief survey of text mining. In *Ldv Forum* (2005), vol. 20, Citeseer, pp. 19–62.

[10] Legany, C., Juhasz, S., and Babos, A. Cluster validity measurement techniques. In *Proceedings of the 5th WSEAS international conference on artificial intelligence, knowledge engineering and data bases* (2006), World Scientific and Engineering Academy and Society (WSEAS) Stevens Point . . . , pp. 388–393.

[11] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. Understanding of internal clustering validation measures. In *2010 IEEE international conference on data mining* (2010), IEEE, pp. 911–916.

[12] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representation in vector space. arXiv preprint arXiv:1301.3781 (2013).

[13] Mostafa, K., Attalla, A., and Hegazy, T. Data mining of school inspection reports to identify the assets with top renewal priority. *Journal of Building Engineering 41* (2021), 102404.

[14] Murtagh, F., and Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2*, 1 (2012), 86–97.

[15] Patel, E., and Kushwaha, D. S. Clustering cloud workloads: K-means vs gaussian mixture model. *Procedia Computer Science 171* (2020), 158–167.

[16] Rejito, J., Atthariq, A., and Abdullah, A. Application of text mining employing k-means algorithms for clustering tweets of tokopedia. In *Journal of Physics: Conference Series* (2021), vol. 1722, IOP Publishing, p. 012019.

[17] Rendón, E., Abundez, I. M., Gutierrez, C., Zagal, S. D., Arizmendi, A., Quiroz, E. M., and Arzate, H. E. A comparison of internal and external cluster validation indexes. In *Proceedings of the 2011 American Conference, San Francisco, CA, USA* (2011), vol. 29, pp. 1–10. 1

[18] Velmurugan, T., and Santhanam, T. Computational complexity between k-means and kmedoids clustering algorithms for normal and uniform distributions of data points. *Journal of computer science 6*, 3 (2010), 363.

[19] Vijayarani, S., Ilamathi, M. J., Nithya, M., et al. Preprocessing techniques for text miningan overview. *International Journal of Computer Science & Communication Networks 5*, 1 (2015), 7–16.

Appendix A: Ground truth based on “Crisis organization” topic.

Cluster	Documents		
4	make sure crisis organisation member well trained via resilience academy		
11	build crisis organisation	Number of clusters	Cluster Group
11	include b2b crisis organisation	1	4
11	include b2b crisis organisation	7	11
11	include b2b crisis organisation	2	25
11	include b2b crisis organisation	9	27
11	include b2b crisis organisation	4	31
11	include b2b crisis organisation		
25	make network administration available crisis organisation		
25	make network administration available crisis organisation		
27	involve crisis organisation		
27	make crisis organisation member aware crisis tool available phone sims apps ncv		
27	make crisis organisation member aware crisis tool available phone sims apps ncv		
27	make crisis organisation member aware crisis tool available phone sims apps		
27	make crisis organisation member aware crisis tool available phone sims apps		
27	make crisis organisation member aware crisis tool available phone sims apps		
27	make crisis organisation member aware crisis tool available phone sims apps		
27	make crisis organisation member aware crisis tool available phone sims apps		
27	make crisis organisation member aware crisis tool available phone sims apps		
31	establish crisis organisation right mandate procedure resource skill		
31	establish crisis organisation right mandate procedure resource skill		
31	establish crisis organisation right mandate procedure resource skill		
31	train staff crisis organisation		

(a). Crisis organization cluster on GMM

Cluster	Docs		
9	involve crisis organisation		
9	make crisis organisation member aware crisis tool available phone sims apps ncv	Number of cluster	Cluster group
9	make crisis organisation member aware crisis tool available phone sims apps ncv	11	9
9	make crisis organisation member aware crisis tool available phone sims apps	7	18
9	make crisis organisation member aware crisis tool available phone sims apps	1	19
9	make crisis organisation member aware crisis tool available phone sims apps	4	34
9	make crisis organisation member aware crisis tool available phone sims apps		
9	make crisis organisation member aware crisis tool available phone sims apps		
9	make crisis organisation member aware crisis tool available phone sims apps		
9	make network administration available crisis organisation		
9	make network administration available crisis organisation		
18	build crisis organisation		
18	include b2b crisis organisation		
18	include b2b crisis organisation		
18	include b2b crisis organisation		
18	include b2b crisis organisation		
18	include b2b crisis organisation		
18	include b2b crisis organisation		
19	make sure crisis organisation member well trained via resilience academy		
34	establish crisis organisation right mandate procedure resource skill		
34	establish crisis organisation right mandate procedure resource skill		
34	establish crisis organisation right mandate procedure resource skill		
34	train staff crisis organisation		

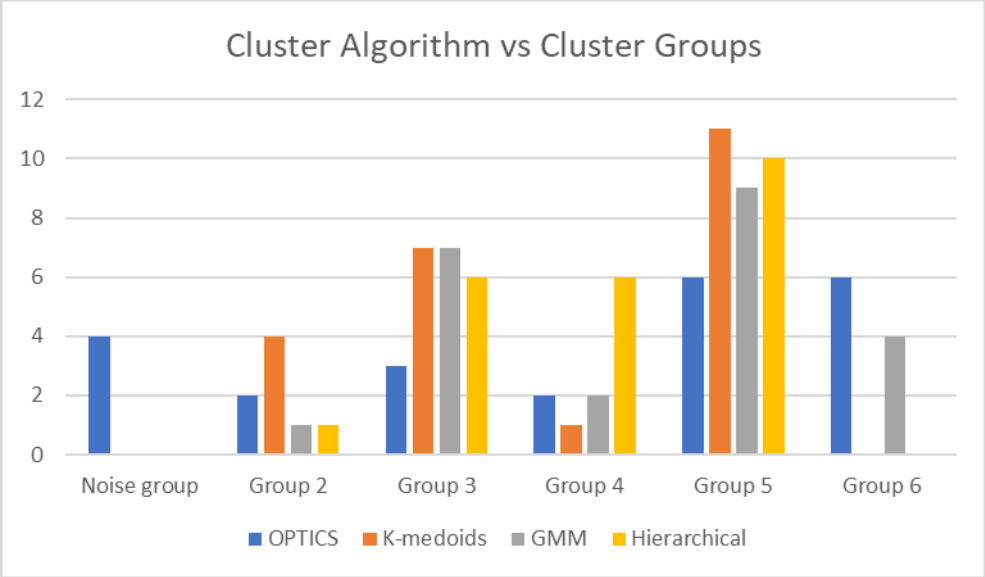
(b). Crisis organization cluster on K-medoids

Cluster	Docs		
-1	build crisis organisation		
-1	involve crisis organisation	Number of clusters	Cluster group
-1	make sure crisis organisation member well trained via resilience academy	4	Noise
-1	train staff crisis organisation	2	5
5	make crisis organisation member aware crisis tool available phone sims apps ncv	6	6
5	make crisis organisation member aware crisis tool available phone sims apps ncv	2	7
6	make crisis organisation member aware crisis tool available phone sims apps	3	24
6	make crisis organisation member aware crisis tool available phone sims apps	6	31
6	make crisis organisation member aware crisis tool available phone sims apps		
6	make crisis organisation member aware crisis tool available phone sims apps		
6	make crisis organisation member aware crisis tool available phone sims apps		
6	make crisis organisation member aware crisis tool available phone sims apps		
7	make network administration available crisis organisation		
7	make network administration available crisis organisation		
24	establish crisis organisation right mandate procedure resource skill		
24	establish crisis organisation right mandate procedure resource skill		
24	establish crisis organisation right mandate procedure resource skill		
31	include b2b crisis organisation		
31	include b2b crisis organisation		
31	include b2b crisis organisation		
31	include b2b crisis organisation		
31	include b2b crisis organisation		
31	include b2b crisis organisation		

(c). Crisis organization cluster on OPTICS

Cluster	Documents		
4	make sure crisis organisation member well trained via resilience academy		
7	build crisis organisation	Number of clusters	Cluster group
7	establish crisis organisation right mandate procedure resource skill	1	4
7	establish crisis organisation right mandate procedure resource skill	6	7
7	establish crisis organisation right mandate procedure resource skill	6	25
7	involve crisis organisation	10	34
7	train staff crisis organisation		
25	include b2b crisis organisation		
25	include b2b crisis organisation		
25	include b2b crisis organisation		
25	include b2b crisis organisation		
25	include b2b crisis organisation		
25	include b2b crisis organisation		
34	make crisis organisation member aware crisis tool available phone sims apps ncv		
34	make crisis organisation member aware crisis tool available phone sims apps ncv		
34	make crisis organisation member aware crisis tool available phone sims apps		
34	make crisis organisation member aware crisis tool available phone sims apps		
34	make crisis organisation member aware crisis tool available phone sims apps		
34	make crisis organisation member aware crisis tool available phone sims apps		
34	make crisis organisation member aware crisis tool available phone sims apps		
34	make crisis organisation member aware crisis tool available phone sims apps		
34	make network administration available crisis organisation		
34	make network administration available crisis organisation		

(d). Crisis organization cluster on Hierarchical Agglomerative


Cluster Algorithm vs Cluster Groups
12
10
8
6
4
2
0
Noise group
Group 2
Group 3
Group 4
Group 5
Group 6
OPTICS
K-medoids
GMM
Hierarchical