KURAWAL Jurnal Teknologi, Informasi dan Industri *Volume 3 Nomor 2 - Oktober 2020*
***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal

IMPLEMENTASI *TEXT MINING* DALAM PENGELOMPOKAN DATA TWEET PERTANIAN INDONESIA DENGAN K-MEANS

Hafiz Irsyad [1], M Rizky Pribadi [2]

Teknik Informatika, STMIK Global Informatika MDP, Jl. Rajawali No. 14, Palembang, Indonesia
email : hafizirsyad@mdp.ac.id [1], rizky@mdp.ac.id [2]

Abstrak

Pertanian 4.0 merupakan suatu gebrakan dimana konsumen lebih dekat pada petani atau para perusahaan pertanian. Salah satu bentuk pertanian 4.0 ini adalah pertanian digital agar setiap kegiatan pertanian dapat terekam, menghasilkan data dan informasi terhadap bentuk dukungan untuk aktivitas pertanian di Indonesia. Pada penelitian ini menerapkan text mining pada data tweet agar dapat mengelompokkan data tersebut dengan menggunakan Algoritma K-Means. Dalam implementasi peneletian ini dibantu dengan menggunakan 2 tools, yakni orange tools untuk melakukan text processing dan Rapidminer untuk melakukan pengolahan algoritma K-Means. Hasil dari penerapan algortima K-Means terdapat 5 klaster, yaitu Pangan, Produksi, Lahan, Ekspor dan Teknologi. Dari 5 (lima) klaster tersebut kemudian menggunakan operator % performance pada rapidminer untuk mendapatkan rata-rata akurasi terhadap klaster tersebut adalah 0.344%. maka hasil dari penelitian ini terdapat 2 klaster yang nilainya tinggi yaitu kluster 0 Pangan dengan nilai 0.528% dan kluster 2 Produksi dengan nilai 0.523% dan untuk kluster yang nilai paling rendah adalah klaster 3 tentang ekspor dengan nilai 0.123% dengan hasil tersbut artinya implementasi text mining dapat dilakukan pada tools rapidminer.

Kata Kunci :
Pertanian, Indonesia, K-Means, Rapidminer

Abstract

Agriculture 4.0 is a breakthrough where consumers are closer to farmers or agricultural companies. One form of agriculture 4.0 is digital agriculture so that each agricultural activity can be recorded, producing data and information on forms of support for agricultural activities in Indonesia. In this study applying text mining to data tweets in order to group the data using the K-Means algorithm. The implementation of this research is supported by using 2 tools, namely the orange tool for text processing and Rapidminer for processing the K-Means algorithm. The results of the application of the K-means algorithm There are 5 clusters, namely Food, Production, Land, Export and Technology. From these 5 (five) clusters then using the% performance operator on the Quickminer to get an average accuracy of the cluster is 0.344%. Then the results of this study include 2 high-value clusters namely 0 Food clusters with a value of 0.528% and 2 Production clusters with a value of 0.523% and for the lowest value clusters are Clusters 3 about exports with a value of 0.123%. with these results means that the implementation of text mining can be done on rapidminer tools.

***Keywords* :**
Pertanian, Indonesia, K-Means, Rapidminer

1. PENDAHULUAN

Indonesia merupakan negara yang mempunyai letak geografis yang sangat mendukung dalam sector pertanian. Selain sektor pertanian yang juga merupakan roda penggerakan perekonimian Indonesia adalah perdagangan dan industri. Seiring waktu berjalan dari tahun 2017 produksi pertanian Indonesia terus menanjak naik terutama padi, yang sangat mengalami pertumbuhan sebesar 2.56% [1]. dalam hal ini pemerintah sebagai penggerak roda perekonomian dan menyukseskan pertanian 4.0 maka perlu untuk melakukan perluasan bisnis, sehingga dapat mendongkrak perekonomian bangsa terutama pada sektor pertanian.

Twitter merupakan salah satu dari sekian banyak media sosial yang difungsikan sebagai pemasaran produk, dimana Twitter sebagai media sosial yang sangat popular setelah Facebook,

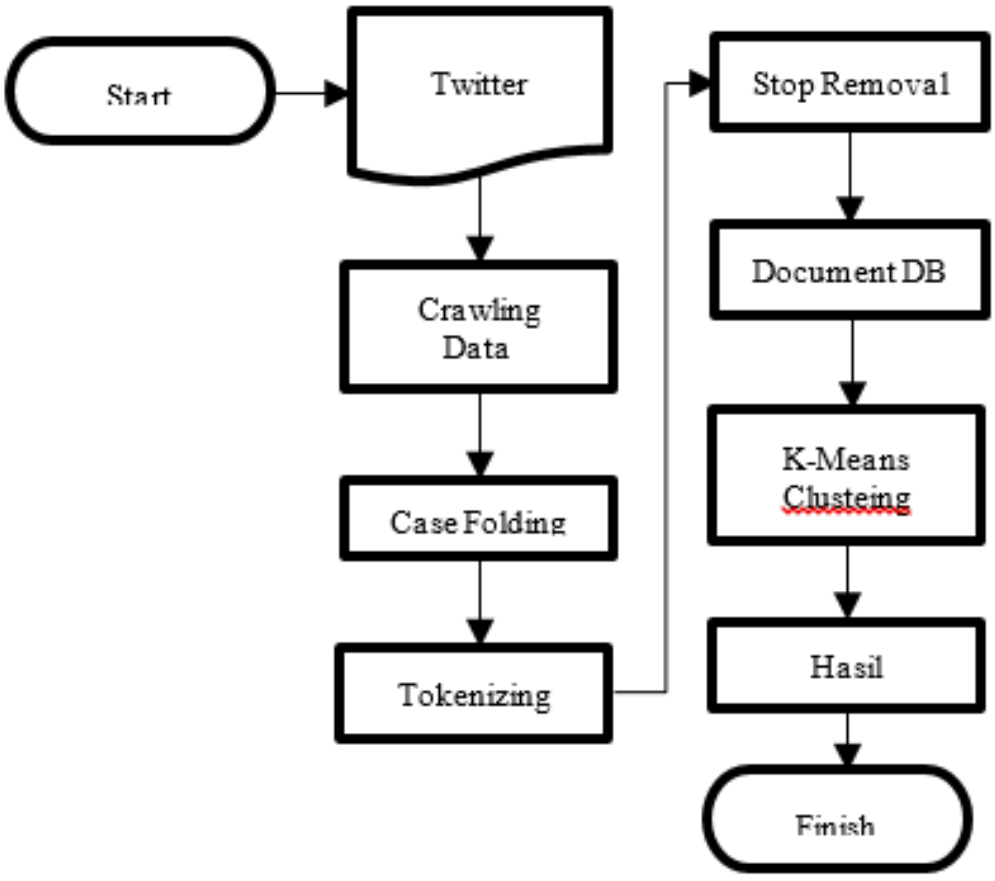
***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal

Instagram dan lain-lain. Pada tahun 2016 indonesia memiliki pengguna aktif sebanyak 24.340 juta orang. Twitter dapat menampung sebanyak 280 karakter, kemudian diolah menjadi sebuah statement [2]. Dewasa ini media sosial Twitter dimanfaatkan sebagai perluasan bisnis, dimana pengguna bisa menemukan pelaku bisnis lain sehingga bisa menjadi teman atau pengikut (*followers*) dan tentu bisa saling berinteraksi.

Pengambilan data *tweet* dapat memanfatkan fasilitas dari Twitter API dan menggunakan *Tools* Orange untuk mengambil data *tweet* tersebut dari Twitter. Untuk mempermudah mengetahui jenis konten dari sejumlah data *tweet*, maka perlu dilakukan proses *Text Mining* terhadap data *tweet* tersebut dengan menerapkan teknik *clustering* [3]. Pada *Text Mining*, teknik *clustering* digunakan untuk mengelompokkan data tekstual berdasarkan kesamaan konten yang dimiliki ke dalam beberapa klaster, sehingga didalam setiap klaster akan berisi data tekstual dengan konten semirip mungkin [4]. Dalam tahapan Teknik clustering ini menggunakan *tools* tambahan agar hasil bisa menjadi maksimal, *tools* yang digunakan adalah Rapidminer.

2. METODE PENELITIAN

Gambaran Umum dalam proses penelitian ini dapat dilihat pada gambar 1. Pada Gambar. 1. Diatas dapat dijelaskan sebabgai berikut:



Gambar 1. Gambaran Umum Penelitian

Tabel 1. Hasil *Crawling* dari Twitter

***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal

No *Tweet*	Teks	Tanggal Tweet	Jumlah Retweet
1	Cara efektif mengendalikan hama, salah satunya menanam refugia. Refugia sebagai tempat berlindung dan sumber makanan bagi musuh alami hama. Jika tanaman refugia banyak Predator hama akan bisa menekan populasi hama… https://t.co/gxeJeknbvX	2019-11-10 02:39:22	4
2	"Ubi jalar 1.000 ha di Karanganyar, Jateng, produktivitas 40-45 ton/ha, harga Rp 3.000-3.500 perkg. Ubi ini kemudian diolah menjadi stik dan keripik. Stik diekspor ke Korea Selatan.	2019-11-06 09:31:30	2
3	Yang menarik setelah musim tanam… https://t.co/FXtzMcbdHp"	2019-11-05 08:04:26	2

1. *Crawling Data* Twitter

   Pengambilan data dilakukan dengan menggunakan *Tools* Orange secara *real time* dalam rentan waktu Januari 2019 sampai dengan Oktober 2019.

2. Case Folding

   Case Folding berarti mengubah semua huruf yang ada pada setiap *tweet* dari huruf besar menjadi huruf kecil.

Tabel 2. Hasil *Case Folding*

No Tweet	Teks hasil *case folding*
1	cara efektif mengendalikan hama, salah satunya menanam refugia. refugia sebagai tempat berlindung dan sumber makanan bagi musuh alami hama. jika tanaman refugia banyak predator hama akan bisa menekan populasi hama
2	ubi jalar 1.000 ha di karanganyar, jateng, produktivitas 40-45 ton/ha, harga rp 3.000-3.500 perkg. ubi ini kemudian diolah menjadi stik dan keripik. stik diekspor ke korea selatan
3	yang menarik setelah musim tanam…

Tabel 3. Hasil *Tekonizing*

***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal

No Tweet	Teks hasil case folding
1	cara efektif mengendalikan hamasalah satunya menanam refugia-refugia sebagai tempat berlindung dan sumber makanan bagi musuh alami hamajika tanaman refugia banyak predator hama akan bisa menekan populasi hama.
2	ubi jalar 1000 ha di karang anyar jateng produktivitas 40-45 ton/ha harga rp 3.000-3.500 perkg ubi ini kemudian diolah menjadi stik dan keripikstik diekspor ke korea selatan
3	yang menarik setelah musim tanam

3. *Tekonizing*

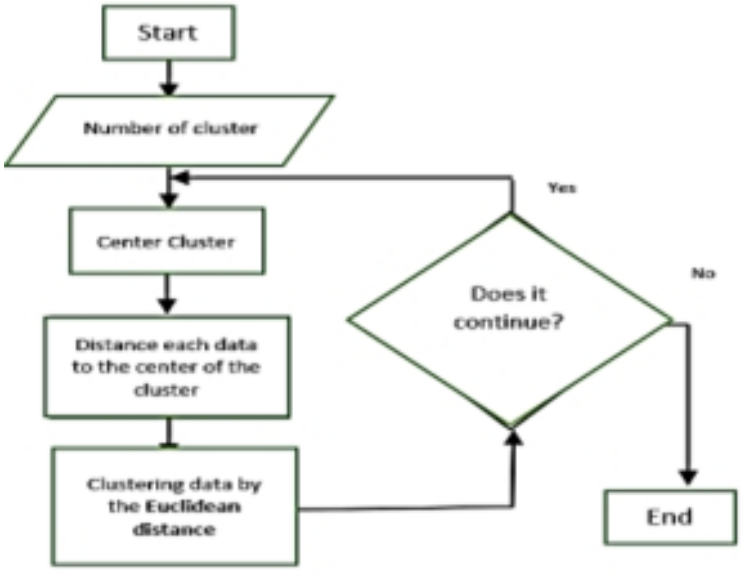
   Tekonizing berfungsi untuk memotong kalimat pada tweet berdasarkan setiap kata.

4. *Stopword Removal*

   Pada proses *stopword removal*, penghilangan kata-kata yang dianggap tidak penting atau tidak menggambarkan isi dari sebuah *tweet*.

Tabel 4. Hasil *Stop Removal*

No *Tweet*	Teks hasil *Tokenizing*	No *Tweet*	Teks hasil *Tokenizing*
1	cara efektif mengendalikan hama salah satunya menanam refugia-refugia sebagai tempat berlindung sumber makanan musuh alami hama tanaman refugia predator hama menekan populasi hama	2	ubi jalar 1000 hakaranganyar jateng produktivitas 40-45 ton/ha harga rupiah 3.000-3.500 perkg ubi diolah menjadi stik keripik stik diekspor korea selatan



Gambar 2. *K-Means Clustering*

5. *K-Means Clustering*

***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal

K-means clustering adalah Teknik yang paling tepat untuk mendapatkan kelompok dari topik penelitian yang ada. *K-means clustering* kemudian digunakan dalam penelitian untuk membantu mengalisis sebuah kelompok. Algoritma *K-means clustering* dimulai dengan menentukan jumlah *cluster* yang akan dibentuk [1]. Proses K-Means *Clustering* dapat kita lihat pada gambar 2 dibawah ini [6].

3. HASIL DAN PEMBAHASAN

3.1. Tahapan Proses Data

Implementasi *data mining* pada pertanian Indonesia menggunakan metode *K-Means* dengan menggunakan 5 penilaian yang terkait kata yang sering muncul saat melakukan *tweet*. Data tweet yang sering muncul. Data yang sering muncul dapat kita lakukan dengan melakukan *crawling data* dengan *tools* Orange, terdapat 350 tweet dari para pengguna twitter yang kemudian disusun menjadi sebanyak 1336 kata. Seluruh kata tersebut dapat kita lihat pada *worcloud* yang telah ditarik dari *tools* Orange.

Berdasarkan hasil dari *worcloud* tersebut dapat kita proses menjadi *variable* dari setiap *tweet*, dengan komponen matriks berupa jumlah dari suatu kata yang ada pada setiap *tweet*. Hasil tersebut dapat dilihat pada tabel 5.

Gambar 3. Hasil *Wordcloud*

Tabel 5. Hasil Proses teks

No *Tweet*	Pangan	Produksi	Lahan	Ekspor	Teknologi
1	8	1	1	1	1
2	7	5	3	3	0
…	3	7	2	2	0
3	4	3	3	3	0

Tabel 6. *Centroid* Data

***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal

Data Cluster	Nilai
Pangan	523
Produksi	226
Lahan	528
Ekspor	123
Teknologi	262

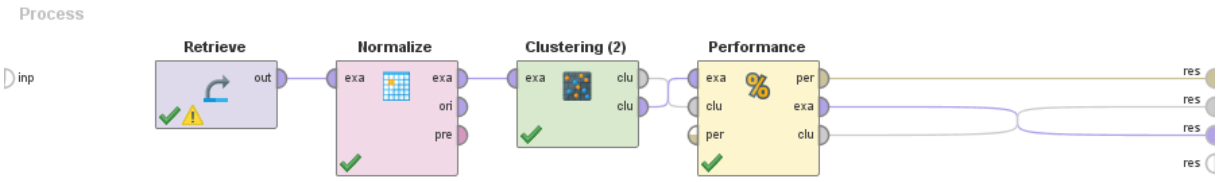
3.2. *Centroid Data*

Pengelompokan data pertanian Indonesia pada twitter dengan menggunakan tools rapidminer studio, maka dari 5 kluster tersebut diperoleh nilai *centroid*-nya. Penentuan nilai centroid awal dilakukan dengan menentukan nilai terbesar sampai dengan nilai terendah. Hasil tersebut dapat dilihat pada tabel 6.

3.3. Implementasi *tools* Rapidminer

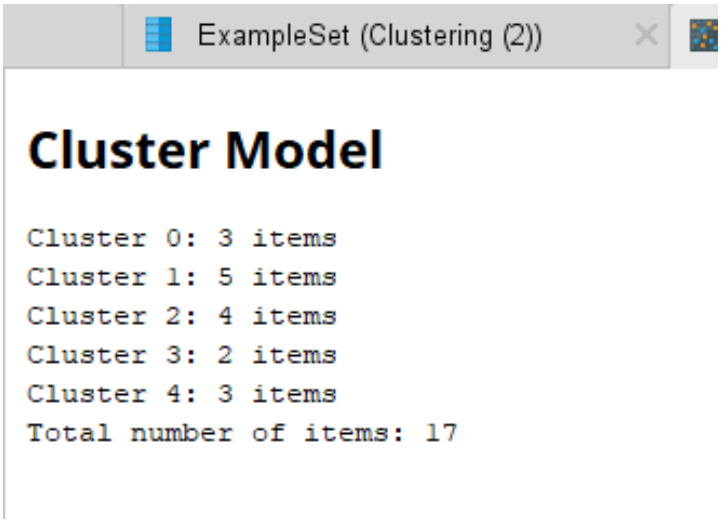
Pada tahap pengelompokan data pertanian, awalnya menggunakan *tools* orange untuk mengambil atau mengelompokan teks yang sering digunakan para netizen di twitter, pada tahapan implementasi penerapan algoritma K-Means penulis menggunakan *tools* rapidminer studio, data awal teks yang sering digunakan dikonversikan kedalam excel dan di*import* ke rapidminer. Algoritma k-means mengelompokan data berdasarkan atribut pada jarak pusat kluster yang kemudian membentuk data seperti pada tabel 2. Proses itertasi pada eksekusi k-means untuk mengelompokkan data berdasarkan pusat kluster terhadap nilai jarak. Nilai jarak pusat kluster akan terus berubah menjadi nol hingga pengelompokan data sama dengan kluster dan iterasi sebelumnya [7].

Pada gambar 4. Dapat dijelaskan bahwasan data yang dibaca dengan menggunakan data excel, yang mana data excel ini adalah hasil dari tools orange yang melakukan proses pemilihan teks yang sering digunakan. Setelah data excel digunakan maka dapat menerapkan normalisasi untuk memperoleh hasil dari excel tersebut. Tahapan berikut menggunakan *clustering* agar dapat mengklasifikasikan yang telah dikelompokan pada excel sebelumnya. Hasil pengelompokan akhir dapat dilihat pada gambar 5.



Gambar 4. Design K-Means pada Rapidminer Studio

***e-issn* : 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal



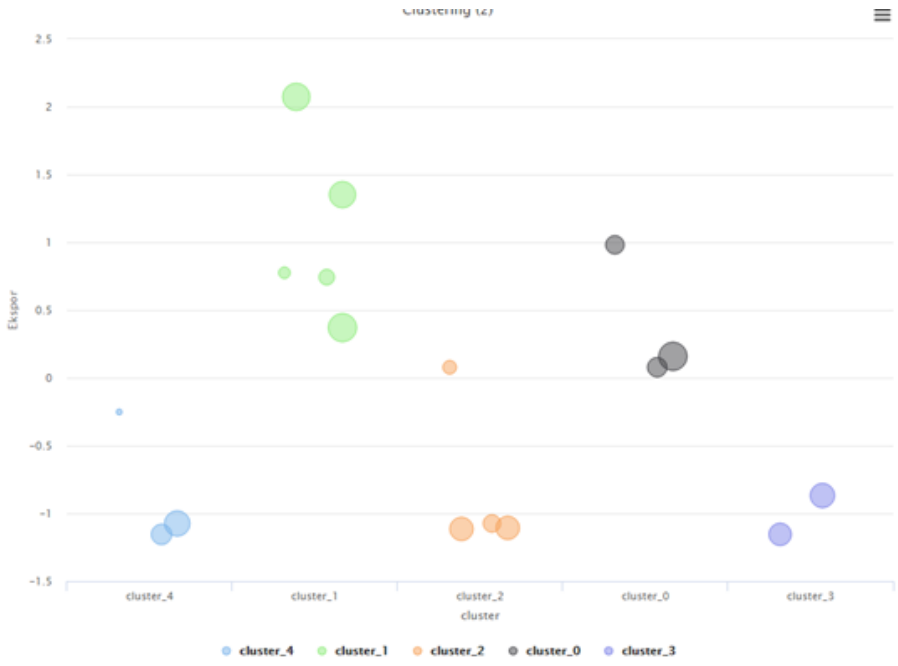
Gambar 5. Hasil kluster

Pada gambar 5 telah didapatkan hasil kluster maka dapat dilihat hasil pada gambar 6 .

Pada gambar 6 nilai cluster akan terus berubah berdasarkan jumlah iterasi yang diterapkan. Ini dikarenakan jarak setiap data ke setiap *centroid* dengan data yang ada pada pusat *cluster* berbeda pada setiap iterasi.

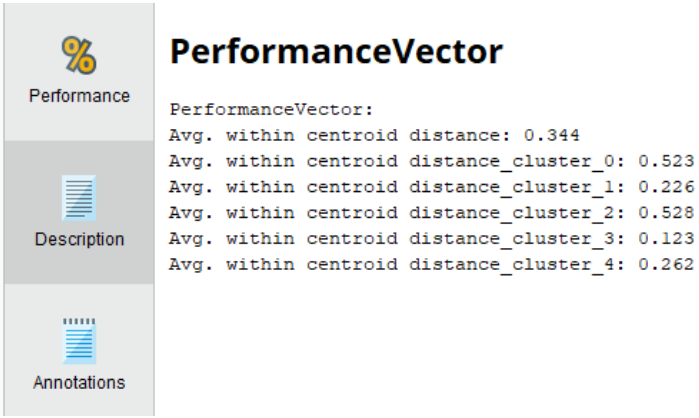
Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Pangan	-0.940	0.555	0.091	-0.682	0.349
Produksi	-0.282	0.123	1.007	-0.097	-1.202
Lahan	-0.345	0.667	-1.011	1.387	-0.345
Ekspor	0.459	1.117	-0.671	-0.979	-0.773
Teknologi	1.872	-0.503	-0.503	-0.503	-0.028

Gambar 6. Hasil final nilai *centroid*



Gambar 7. Hasil Klustering dengan Rapidminer Studio

***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal



Gambar 8. Akurasi hasil Klustering operator % *Performance*

Setelah mendapatkan hasil dari klustering dengan *operator clustering* pada rapidminer studio, langkah selanjutnya adalah menggunakan satu *operator* untuk mengukur akurasi dari K-Means tersebut, operator tersebut adalah % *performance*. % *performance* digunakan untuk melakukan evaluasi kinerja dari metode operator *clustering* yang berbasiskan centroid. % *performance* yang dapat dihasilkan dari operator tersebut adalah Avg. *within centroid distance* per setiap kluster yang telah ditentukan. Hasil dari kinerja operator % *performance* yang ada pada rapidminer dapat dilihat pada gambar 8.

4. KESIMPULAN

Pengelompokan data *tweet* pertanian dapat dilakukan dengan data *mining*. Metode data mining yang digunakan adalah k-means dengan memanfaatkan *Tools* Rapidminer, sedangkan untuk pengelompokan hasil *tweet* dengan memanfaatkan Tools Orange. Adapun hasil yang dapat diperoleh dari klustering dengan menggunakan algoritma K-Means adalah:

1. Penerapan algoritma proses *Text Mining* untuk melakukan *clustering* dengan metode K-means pada data *tweet* pertanian Indonesia menghasilkan sejumlah 5 klaster tweet.
2. Berdasarkan hasil proses penentuan jenis konten dan perhitungan rata-rata jumlah *retweet* pada tiap klaster, didapatkan bahwa jenis konten pada klaster yang memiliki jumlah *retweet* yang tinggi diantaranya Klaster 2 yaitu pangan dan klaster 0 produksi.
3. Berdasarkan hasil proses dari rata-rata *retweet* pada setiap klaster yang paling rendah adalah klaster 3 yaitu teknologi.

5. REFERENSI

[1] B. P. Statistik, "www.bps.go.id," 5 2 2018. [Online]. Available: https://www.bps.go.id/pressrelease/2018/02/05/1519/ekonomi-indonesia-triwulan-iv-2017--tumbuh-5-19-persen.html. [Accessed 10 8 2018].

[2] M. Rani and A. J, "Twitter Data Predicting Geolocation Using Data Mining Techniques," *International Journal of Innovative Research in Computer,* vol. 4, no. 6, p. 10446, 2016.

[3] M. S. Kini, Devi, D. P.G and N. Chiplunkar, "Text mining Approach to Classify Technical Research Document using Naïve Bayes," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 4, no. 7, pp. 386-391, 2015.

[4] E. Yulian, "Text Mining dengan K-Means Clustering pada Tema LGBT dalam Arsip Tweet Masyarakat Kota Bandung," *JURNAL MATEMATIKA "MANTIK",* vol. 4, no. 1, pp. 53-58, 2018.

***e-issn :* 2615-6474**
p-issn : 2620-3804
URL : https://jurnal.machung.ac.id/index.php/kurawal

[5] Hafiz I, . M. Rizky. , "Klasifikasi Opini Masyarakat Terhadap Jasa ISP MYRepublic dengan Naïve Bayes," Jurnal JNTETI, vol. 8, no. 1, 2019.

[6] Srihari. [Online]. Available: https://cedar.buffalo.edu/~srihari/CSE626/Lecture-Slides/. [Accessed 11 08 2018].

[7] Sudirman, W. P. Agus and W. Anjar, "Data Mining Tools | rapidminer: K-Means Method on Clustering of rice crops by Province as Efforts to Stabilize Food Crops in Indonesia, " Nommensen International Conference on Technology and Engineering, 2018.