p-ISSN 2087-1716
e-ISSN 2548-7779

ILKOM Jurnal Ilmiah Volume 10 Nomor 2 Agustus 2018

ANALISA CLUSTERING PHISING DENGAN K-MEANS DALAM MENINGKATKAN KEAMANAN KOMPUTER

Suhardi Rustam
Suhardirstm@gmail.com
Universitas Ichsan Gorontalo

Abstrak

Hampir setiap aksi kejahatan di siber merupakan suatu kondisi mengenai aktivitas kriminal dengan menggunakan komputer atau jaringan komputer sebagai alat bantu dan juga sebagai target. Penipuan di website akademik yang paling memiliki resiko. Aksi Phishing ini semakin marak terjadi.Tercatat secara global, jumlah penipuan bermodus phising 42% dari modus selain phising yang dinyatakan dalam website Anti-Phishing Working Group (APWG) dalam laporan bulanannya, mencatat ada 12.845 e-mail baru dan unik serta 2.560 situs palsu yang digunakan sebagai sarana phishing, Selain terjadi peningkatan kuantitas, kualitas serangan pun juga mengalami kenaikan, perlu adanya usaha yang dilakukan oleh para administrator jaringan dalam meningkatkan pengawasan dalam memonitoring aktivitas di jaringan, dalam aksi pencurian data akan melakukan aksi memanipulasi seseorang dengan tampilan situs web tertentu. Dalam penelitian ini sekumpulan dataset akan diklasterisasi menggunakan k-means, algoritma K-Means akan mengelompokan dataset, menghasilkan identifikasi pishing yang lebih akurat dan bermutu.dengan hasil penelitian ini iterasi=10, K-Fold=2 nilai index davis bouldin = 0.119

Kata Kunci: *Klustering, Phising, K-Means,* Keamanan Komputer, Data Mining

Abstract

Almost the crime in cyber is a condition of criminal activity using computers or computer networks as tools and also as a target. Fraud in academic websites the most at risk. The action of Phishing is on the rise. Recorded globally, the number of fraudulent mode phishing 42% of the mode in addition to phishing which is stated in the website Anti-Phishing Working Group (APWG) in its monthly report, noting there 12.845 e-mail new and unique as well as 2.560 a fake site that is used as a means of phishing, in Addition to increase the quantity, the quality of the attacks is also increasing, the need for the work done by the network administrator in improving surveillance in monitoring activity on the network, in the action of data theft will perform the action of manipulating someone with the appearance of a particular web site. In this study a set of datasets will be clustering using k-means, K-Means algorithm will classify the dataset, resulted in the identification of phishing that is accurate and certifiable. With the results of this research iteration=10, the K-Fold=2 the of the Bouldin Davis index = 0.119.

Keywords: Clustering, Phishing, K-Means Clustering, Computer Security, Data Mining

1. Pendahuluan

Hal yang merisaukan dari perkembangan teknologi informasi yang senantiasa berubah serta cepatnya dari perkembangan software, keamanan merupakan suatu isu yang sangat krusial dan setiap orang mempertaruhkan waktu dan biaya untuk melindungi data privasi di internet[1], mahalnya biaya perlindungan data baik itu perlindungan fisik, data maupun aplikasi. Hampir seluruh aktivitas sehari-hari di tuangkan ke dalam internet, interaksi dengan internet dengan memberikan data pribadi pada saat memiliki akun tertentu yang terhubung ke data global. dan Masalah keamanan dan kerahasiaan data merupakan salah satu aspek penting dari suatu sistem informasi. Keamanan komputer adalah suatu cabang teknologi yang dikenal dengan nama keamanan informasi yang diterapkan pada komputer. Sasaran keamanan komputer antara lain adalah sebagai perlindungan informasi terhadap pencurian atau korupsi, atau pemeliharaan ketersediaan.

Aksi Phishing ini semakin marak terjadi.Tercatat secara global, jumlah penipuan bermodus phising 42% dari modus selain phising yang dinyatakan dalam website Anti-Phishing Working Group (APWG) dalam laporan bulanannya, mencatat ada 12.845 e-mail baru dan unik serta 2.560 situs palsu yang digunakan sebagai sarana phishing.Selain terjadi peningkatan kuantitas, kualitas serangan pun juga mengalami kenaikan. Artinya, situs-situs palsu itu ditempatkan pada server yang tidak menggunakan protokol standar sehingga terhindar dari pendeteksian, perlu adanya usaha yang dilakukan oleh para administrator jaringan dalam meningkatkan pengawasan lalu lintas data di jaringan komputer dengan berbagai cara. Pada penelitian sebelumnya mengatasi ancaman penipuan dengan menonaktifkan

layanan-layanan komputer di jaringan, penelitian sebelumnya analisa phising hanya membahasa artribut *SSL final_state* yang dipercaya dan tidak dipercaya oleh penyedia ternama[2].

Penelitian ini akan membahas bagaimana mengidentifikasi Pishing dengan k-means Klustering dan Bagaimana menerapkan Algoritma K-Means Klustering, perlu dibuatkan analisa yang akan menjadi rujukan bagi administrator, pengawas dan perencana jaringan dan system informasi. Selain itu *Clustering* merupakan metode yang digunakan dalam data mining yang cara kerjanya mencari dan menglompokkan data mempunyai kemiripan karakteristik antara data satu dengan data lainnya yang telah diperoleh. Ciri khas dari teknik data mining[3]. Metode *clustering* yang mempunyai sifat efesien dan cepat yang dapat digunakan salah satunya adalah metode k-means, metode ini bertujuan untuk membuat *cluster* objek berdasarkan atribut menjadi *k* partisi.cara kerja metode ini adalah mula – mula ditentukan *cluster* yang akan dibentuk, pada elemen pertama dalam tiap *cluster* dapat dipilih untuk dijadikan sebagai titik tengah (*centroid*), selanjutnya akan dilakukan pengulangan langkah –langkah hingga tidak ada objek yang dapat dipindahkan lagi[4].

kluster merupakan suatu teknik klasifikasi tanpa pengawasan, menurut kriteria kesamaan tertentu untuk mengklasifikasikan dataset, sehingga objek dari kelas yang mungkin sama, tapi itu adalah keragaman mungkin antara objek yang berbeda, Fungsi yang paling penting dari algoritma klustering mengukur kesamaan untuk menentukan bagaimana menutup dua pola data ke satu sama lain.

Pada algoritma k-means, dikenal membutuhkan parameter input sebanyak *k, parameter input k sebagai atribut* dan membagi sekumpulan *n,* dimana n adadalh jumlah seluruh atribut objek kedalam *parameter input cluster* sehingga tingkat kemiripan antar anggota dalam satu *cluster* tinggi sedangkan tingkat kemiripan atribut dengan anggota pada *cluster* lain sangat rendah. Kemiripan atribut terhadap *cluster* diukur dengan kedekatan objek terhadap nilai *mean* pada *cluster* atau dapat disebut sebagai *centroid cluster* atau pusat massa, jika tingkat kemiripan tersebut tinggi maka atribut tersebut dinyatakan sebagait atribut phising dan diberikan saran untuk memonitoring atribut tersebut agar segera diberikan penanganan. Dengan adanya penelitian ini diharapkan pengetahuan administrator, pengawas, pengguna dan perencana jaringan computer semakin meningkat sehingga keamanan jaringan computer dapat ditingkatkan.

2. Metode

Metode yang digunakan dalam penelitian ini adalah metode data sampel, alasan menggunakan metode ini adalah bahwa pengambilan dataset dari hasil pencarian dan studi dataset dari kumpulan dataset UCI yang terbaik. Metode ini sangat sesuai diterapkan dalam penelitian ini karena sebelum uji coba dataset tersebut di preprosessing untuk menghasilkan dataset yang valid sebelum siap diolah dalam program rapidminer 5, RapidMiner sebagai *tools data mining* memiliki antarmuka yang nyaman, dimana analisa dikonfigurasi dalam sebuah *process view*. Dalam konsep modular untuk *process view* ini, setiap langkah analisis digambarkan dengan sebuah operator dalam proses analisis. Operator-operator ini memiliki *port* untuk *input* dan *output* dimana operator tersebut dapat berkomunikasi dengan operator lain untuk mendapatkan *input data* atau mengirim data yang telah diubah dan menggenerasi model. *Davies-Bouldin Index* merupakan salah satu metode yang digunakan untuk mengukur validitas *cluster* pada suatu metode pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap titik pusat *cluster* dari *cluster* yang diikuti, Evaluasi menggunakan *Davies Bouldin Index* ini memiliki skema evaluasi internal *cluster*, dimana baik atau tidaknya hasil *cluster* dilihat dari kuantitas dan kedekatan antar data hasil *cluster[5],*dapun skema dalam menggunakan metode data sampel adalah sebagai berikut [6].

Gambar 1 skema metode data sampel

2.1 Diagnosa Data Sampel

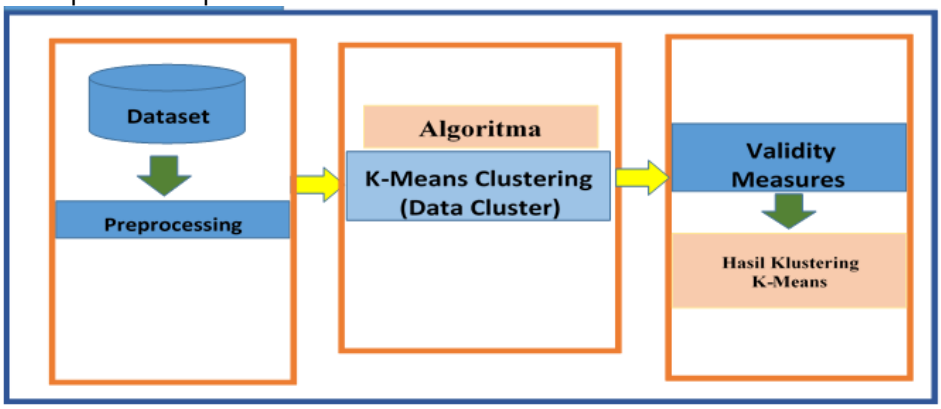
Dalam tahapan ini dimulai dari Identifikasi masalah merupakan proses awal bagaimana memahami persoalan yang ada, seringnya/meningkatnya gangguan/ancaman penipuan melalui jaringan internet, hal ini yang menjadi persoalan bagi administrator/pengawas/perencana jaringan komputer saat ini, selanjutnya pada penentuan metode penelitian yang dilakukan dalam penelitian saat ini adalah metode data sampel, metode ini melakukan pengujian terhadap sejumlah data atau keseluruhan data sampel, pembersihan atribut/preprosessing juga di gunakan untuk menggurangi nilai atribut yang tidak valid dalam data sampel , sehingga metode ini layak digunakan, juga tahap lain yaitu perumusan masalah dalam penelitian ini merumuskan masalah yaitu bagaimana menerapkan algoritma k-means dalam menganalisa clustering phising, adapun tujuan penelitian ini adalah mampu mengklastering data phising dengan metode k-means.

2.2 Perencanaan Data Sampel

Untuk indentifikasi variabel data sampel dalam perencanaan ini yaitu menentukan variabel-variabel yang memiliki kaitan yang dekat antar variabel dan kemudian dari data sampel akan ditentukan besaran jumlah data dari 500 s/d 1000 record data.

2.2.3 Penentuan alur

Berikut ini merupakan alur penelitian



Gambar 2 alur penelitian

2.3 Pengolahan Data Sampel

Penting diperhatikan dalam pengolahan data sampel yaitu pembersihan data atau yang sangat dikenal preprosessing Dataset yang akan diproses terlebih dahulu dilakukan preprosessing untuk membersihkan atribut yang tidak berkaitan atau nilai dari atribut yang tidak valid, selanjutnya adalah pengolahan data dimana pengolah data merupakan pengecekan kembali nilai dan atribut data sehingga bisa diolah dengan tool rapidminer 5, kesiapan tahap inilah yang harus diperhatikan sebelum pembentukan prototype, Prototype merupakan gambar tampilan yang akan dibuat.

2.4 Pengujian data

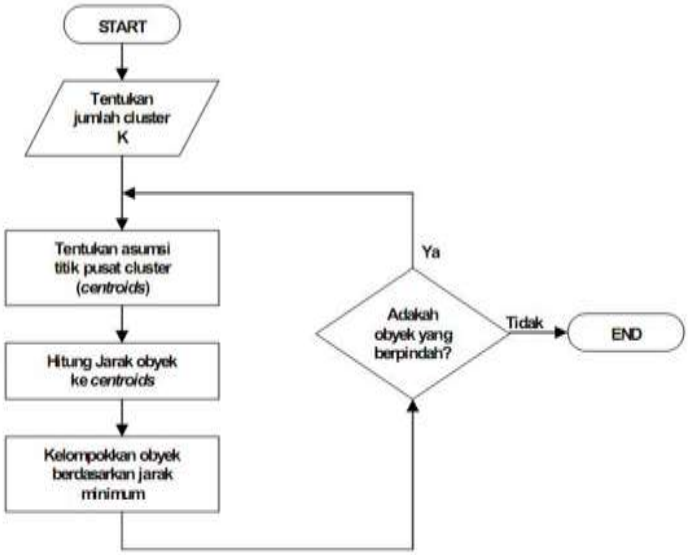
Pengujian data dengan menggunakan tool rapidminer 5 secara berulang dengan menggunakan parameter yang disediakan rapidminer 5 sampai menghasilkan nilai maksimal.

3. Hasil dan Pembahasan

3.1 Implementasi Algoritma K-Means

Dapat dilihat pada gambar 3 di bawah merupakan diagram alur dari metode k-means yang digunakan dalam pengelompokkan pishing, pada umumnya kinerja metode k-means secara berurutan adalah sebagai berikut :

1. Menentukan banyaknya cluster (k)
2. Menentukan centroid
3. Apakah nilai centroidnya berubah? a.Jika ya, hitung jarak data dari centroid b.Jika tidak, selesai.
4. Mengelompokkan data berdasarkan jarak terdekat

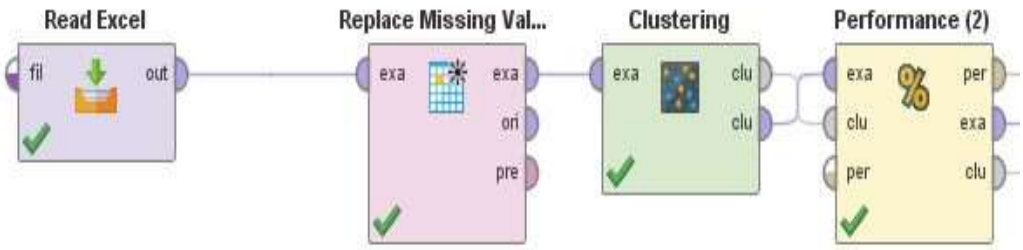


Gambar 3 Alur Implementasi Algoritma K-Means

3.2 Preprocessing

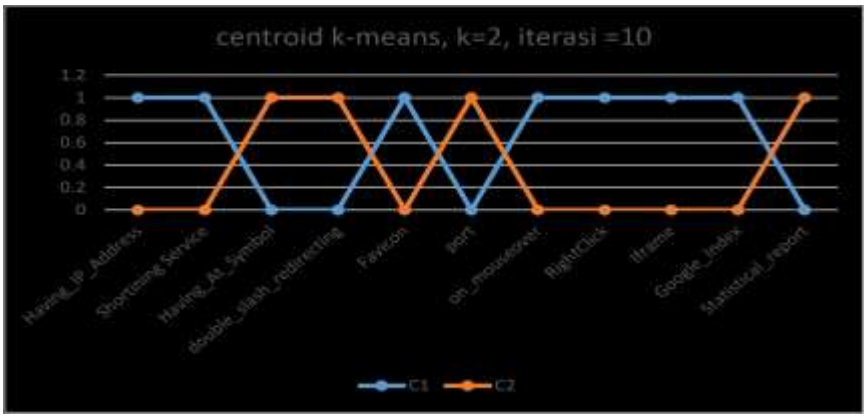
Preprocessing Pada sebuah penelitian data mining terdapat data yang akan diolah dengan metode yang telah ditentukan sebelumnya, pada penelitian ini data yang digunakan adalah dataset pishing yang akan diolah menggunakan metode k-means untuk mengelompokkan data variable serangan tersebut kedalam kelompok variabel yang "sering" dan "tidak sering" berdasarkan beberapa variabel inputan. Variabel inputan yang digunakan dalam pengelompokkan pishing tersebut adalah *Having_IP_Address, Shortining Service, Having_At_Symbol, double_slash_redirecting, Favicon, port, on_mouseover, RightClick, Iframe, Google_Index* dan *Statistical_report*. Kemudian variabel tersebut akan diolah menggunakan metode k-means yang kemudian menghasilkan output kelompok pishing.

Data penelitian yang sedang dilakukan merupakan data pishing sebanyak 300 data yang akan dikelompokkan ke dalam *cluster* "sering(C1)" dan *cluster* "tidak sering (C2)" pengelompokkan tersebut diolah menggunakan algoritma k- means. Sampel dari dataset pishing dapat dilihat pada tabel 1 di bawah ini.



Gambar 4 Model K-Means pada Rapidminer

Gambar 4 Model K-Means pada Rapidminer, yang di mulai dari pembacaan dataset yang dimasukkan melalui Microsoft excel yang telah diklasifikasi variable dan menghilangkan *outlier* gangguan nilai untuk memudahkan pembacaan nilai *numeric*, *Replace Missing Value* merupakan operator untuk membersihkan nilai data ke dalam tipe data *numeric*, operator *Clustering* merupakan model operator untuk metode K-Means untuk menghitung jarak dari setiap kluster data, operator *Performance* untuk pengukur validasi dengan menggunakan metode index davis bouldin.



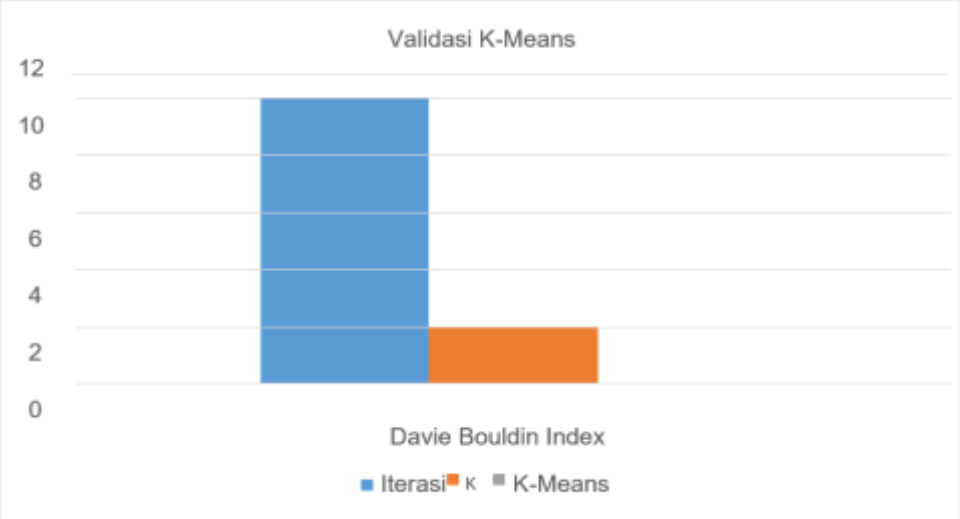
Gambar 5 Grafik Centroid K-Means Klustering

3.4 Analisis dan Validasi Hasil

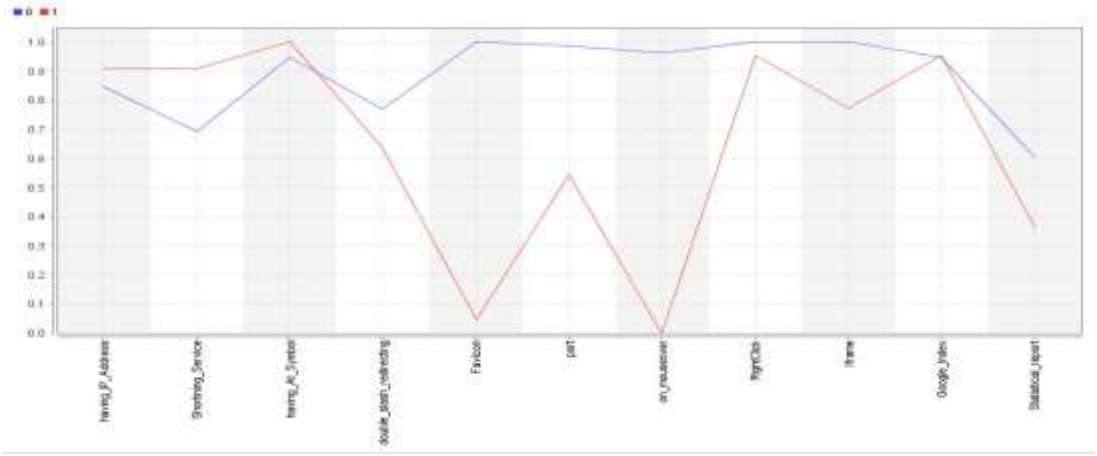
Tabel 1 Hasil Validasi K-Means

Validasi	Iterasi	K	K-Means
Davie Bouldin Index	2	2	2.021
	3	2	1.915
	4	2	2.113
	5	2	1.991
	6	2	2.811
	7	2	2.132
	8	2	2.211
	9	2	1.071
	10	2	0.119

Tabel 1, Hasil Validasi Metode K-Means, dengan validasi Index Davis Bouldin sebagai Performance, iterasi (banyaknya dataset di jalankan) 10, dengan K-Fold (pembagi kluster) 2 maka hasil validasi metode k-means adalah 0.119. hal ini juga ditunjukan pada gambar 6 Grafik Validasi K-Means, Dbi *(Davies bouldin index),* kluster yang memiliki nilai Dbi terendah dianggap sebagai kluster yang paling tepat. seperti di bawah ini.



Gambar 6 Grafik Validasi K-Means



Gambar 7 Plot K-Means Klustering

Gambar 7 Plot K-Means Klustering, merupakan hasil klustering dataset pisihing dengan cluster_0 dan cluster_1.

3.5 Hasil Uji Coba

Tabel 2 Hasil Eksperimen Data Pishing

Dataset	Metode	no	Nama Variabel	C1	C2	Iterasi	K-Fold	Index davies Bouldin (IDB)
Phising	K-Means	1	Having_IP_Address	1	0	10	2	0.119
		2	Shortining Service	1	0			
		3	Having_At_Symbol	0	1			
		4	double_slash_redirect ing	0	1			
		5	Favicon	1	0			
		6	port	0	1			
		7	on_mouseover	1	0			
		8	RightClick	1	0			
		9	Iframe	1	0			
		10	Google_Index	1	0			
		11	Statistical_report	0	1			

Tabel 2 hasil Eksperimen Data Pishing, dengan Dataset Pishing, menggunakan metode K-Means, iterasi=10, K-Fold=2 menghasilkan index davis bouldin = 0.119, dengan pengukuran ini maka Dataset Pishing memiliki validasi mendekati dengan 0 maka hasil nilai index davis bouldin = 0.119 memiliki informasi yang lebih baik.

4. Kesimpulan dan Saran

Berdasarkan hasil eksperimen dan pembahasan, maka dengan ini dapat simpulkan bahwa pengujian model dengan menggunakan algoritma k-means untuk data pishing mendapatkan nilai IDB (*davies bouldin index*) = 0.119, k-fold =2, iterasi = 10 sebagai nilai ukur terhadap validasi data pishing. Maka pengujian model dengan menggunakan k-means dengn IDB (*davies bouldin index*) adalah Lebih baik dalam memberikan pemecahan untuk permasalahan identifikasi pishing untuk keamanan computer yang lebih akurat dan bermutu.

5. Terima Kasih

Saya mengucapkan banyak terima kasih atas sharing ilmu dari rekan sesama dosen dan peneliti pemula di fakultas ilmu komputer universitas ichsan gorontalo

Daftar Pustaka

[1] Musriha, Gilang R. 2014. Pengaruh Intensitas Pemakaian Internet Terhadap Penggunaan Internet untuk Berbelanja Online yang Dimoderasi oleh Consumer Innovativeness Di Surabaya
[2] Salim, Tomy. 2017. Data Mining Mengidentifikasi Website Phising Menggunakan Algoritma C.45. Jurnal TAM (Technology Acceptance Model) Volume 8
[3] Arora, P., Deepali, D., dan Varshney, S. 2015. Analysis of K-Means and KMedoids Algorithm For Big Data. International Conference on Information Security & Privacy (ICISP2015), (hal. 507-512). Nagpur,India.
[4] Agusta Yudi.2007. K-Means – Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika Vol. 3
[5] Bates, A. & Kalita, J. 2016. Counting clusters in twitter posts. *Proceedings of the 2nd International Conference on Information Technology for Competitive Strategies*, pp. 85
[6] Hasibuan, A Zaenal .2007. Metodelogi Penelitian Pada Bidang Ilmu Komputer dan Teknologi Informasi. 20 Juli 2018

