ANALISIS KECENDERUNGAN INFORMASI

by Lppm 2022

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

ANALISIS KECENDERUNGAN INFORMASI MENGGUNAKAN ALGORITMA *HIERARCHICAL AGGLOMERATIVE CLUSTERING*

Herny Februariyanti[1], Jati Sasongko Wibowo[2], Dwi Budi Santoso[3], Muji Sukur[4]
[1,2,3,4]Universitas Stikubank Semarang
Jl. Tri Lomba Juang No. 1 Semarang (50241)
e-mail : hernyfeb@edu.unisbank.ac.id

ABSTRAK

Media Sosial merupakan media online dimana dengan mudah para penggunanya dapat ikut berpartisipasi, saling berbagi (sharing) dan juga dapat memmbuat konten atau menciptakan isi. Media online ini contohnya adalah jejaring social, twitter, wiki, forum serta dunia virtual. Twitter merupakan salah satu media social yang seringkali dan paling umum digunakan orang disegala penjuru dunia. Tujuan dari penelitian ini adalah melakukan proses klastering dengan menggunakan algoritma Hierarchical Agglomerative Clustering yang diimplementasikan pada dokumen teks di twitter di akun twitter @unisbank. Data penelitian ini merupakan data tweet (kicauan) yang berisi isue yang menjadi bahan kicauan mahasiswa yang dikumpulkan berdasarkan hashtag tentang Unisbank Semarang. Untuk mengetahui keterkaitan informasi yang satu dengan lainnya digunakan algoritma Eucledian Distance untuk perhitungan jarak dokumen. Hasil perhitungan jarak dokumen dihasilkan jarak terdekat sebesar 1 dan jarak terjauh sebesar 69. Hasil penelitian ini menyajikan klastering data teks twitter dengan Algoritma Hierarchical Agglomeratif dengan mengambil titik tengah 20 menghasilkan klaster sejumlah 7 buah klaster dengan klaster paling besar adalah dengan jarak eucledian sebesar 28.

Kata kunci *: twitter, hierarchical agglomerative, eucledian distance, klaster.*

ABSTRACT

Social media is an online media where users can easily participate, share, create content or give an opinion to certain information or news. Twitter is social media that is often used by people around the world. This research has a purpose to cluster student tweets using a hierarchical agglomerative algorithm on the @unisbank account. The data of this research are tweets collected based on the hastag about Unisbank Semarang. To find out information from one another using an algorithm to calculate the eucledian distance. The result of calculation of document distances taking the center of 20 results in a total of 7 document clusters, with the largest cluster having an eucledian distance of 28.

Keywords : *twitter, hierarchical agglomerative, eucledian distance, cluster.*

1. PENDAHULUAN

Twitter merupakan salah satu media sosial yang seringkali digunakan oleh masyarakat dalam menyampaikan pesan. Selain itu dengan twitter orang bisa menggunakannya sebagai tempat untuk 'curhat' tentang sesuatu hal baik yang bersifat memuji maupun mencela. Twitter juga merupakan media sosial yang sangat populer, pengguna twitter dapat berekspresi tentang opini yang obyektif mengenai topik yang berbeda (Coletta et al., 2014). Dalam ilmu pemrosesan dokumen teks dikenal dengan istilah analisa opini atau seringkali disebut dengan analisa sentimen (*opinion analysis atau sentiment analysis*) yaitu penelitian dengan menganalisis 'curhatan' di media sosial. Sedangkan yang

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

dimaksud dengan opini sendiri adalah pandangan seseorang terkait suatu hal tertentu secara subyektif. Menurut Kuncoro, 2009 dalam penelitian (Prabawati & Dawud, 2019) yang dimaksud dengan opini adalah pemikiran seseorang yang disajikan dalam bentuk tulisan atau pendapat seseorang mengenai berbagai macam fakta. Artikel opini atau opini itu sendiri merupakan tulisan dari seseorang yang dituangkan secara lepas dengan membahas suatu masalah tertentu yang bersifat aktual maupun kontroversial yang memiliki tujuan untuk memberikan informasi, maupun mempengaruhi orang lain serat meyakinkan orang bahkan bisa juga sebatas ingin menghibur pembaca (bersifat recreatif)

Analisa opini bisa dilakukan secara manual, yaitu dengan cara memonitor berita-berita yang ada di media masa. Hanya saja untuk dokumen teks yang ada di twitter tentu cara manual tidak mungkin bisa dilakukan. Hal ini sangat jelas dikarenakan dokumen teks yang ada di twitter ukurannya sangat besar sekali dan setiap waktu dokumen akan mengalir setiap waktu bahkan setiap detik. Seperti hal nya contoh tweet dengan bahasa Indonesia, diperkirakan setiap hari ada aliran data sejumlah 6 juta tweet. Dengan kondisi yang ada maka sangat diperlukan peranan dari pemrosesan teks (*text processing*) yaitu data dapat secara otomatis diproses. Oleh karena itu pada penelitian ini akan dilakukan pengelompokan dokumen teks dari twitter @unisbank dengan menggunakan metode *Hierarchical Agglomerative Clustering*. Tujuan yang ingin dicapai dari penelitian ini adalah untuk mengetahui kecenderungan mahasiswa terhadap pemberitaan dan mengetahui topik yang seringkali muncul di twitter @unisbank.

a. Definisi Teks Mining

Seringkali teks mining disebut sebagai Teks Data Mining atau yang disingkat TDM, orang juga sering menyebut juga dengan istilah *Knowledge Discovery in Text* (KDT). Secara umum teks mining merupakan proses mengekstraksi suatu informasi yang diambil dari dokumen-dokumen teks *unstructured yaitu dokumen teks* yang memiliki format tidak terstruktur. Text mining dapat diartikan juga sebagai sebuah temuan baru dari suatu informasi yang belum diketahui sebelumnya oleh komputer, dimana akan diekstrak informasi dengan cara otomatis dari sumber-sumber yang tdk tersetruktur dengan sumber yang berbeda-beda. Inti dari pemrosesan text mining ini adalah bagaimana menggabungkan suatu informasi yang dapat dilakukan proses ekstraksi dari banyak sumber. (Tan, 2011). Tujuan utama dilakukan kegiatan atau proses text mining ini adalah pada koleksi dokumen yang cukup besar biasanya dilakukan proses *knowledge discovery,* maka kegiatan teks mining ini sangat membantu sekali. Andika (2015) dalam (Susandi & Sholahudin, 2016) menjelaskan bahwa text mining merupakan bidang ilmu multidisipliner, meliputi *information retrieval*, *information extraction*, *text analysis*, *clustering, categorization, database technology, visualization, natural language processing* (NLP), *machinelearning*, dan data mining. Text mining juga merupakan salah satu ilmu tentang *artificial intelligence* atau sering kita kenal dangan aplikasi kecerdasan buatan (Nugroho, 2011). Dengan text mining maka kita dapat menyelesaikan masalah *information overload* dengan memanfaatkan bermacam-macam teknik yang berkaitan dengan bidang ilmu tersebut. Text mining juga merupakan pengembangan ilmu dari data mining atau sering kita sebut sebagai *knowledge -discovery in database*. Data mining merupakan ilmu pengetahuan yang berusaha untuk menemukan pola-pola yang menarik dari suatu data yang tersimpan dalam basis data dengan ukuran yang sangat besar. Dibandingkan dengan data mining, text mining memiliki nilai komersil lebih tinggi. Hal ini dikarenakan sebagian besar penyimpanan informasi secara umum menggunakan format data dengan tipe teks. Text mining juga menggunakan informasi dalam bentuk tidak terstruktur serta pengujiannya dilakukan untuk menemukan struktur serta arti dari suatu teks yang tersembunyi. Perbedaan text mining dan data mining adalah pada sumber data yang akan digunakan dalam proses analisis dan pengujian.

.

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

b. Pembobotan Istilah

Masalah dalam *term weighting* (pembobotan istilah) adalah bagaimana memutuskan bobot tiap istilah yang muncul dalam *query* atau koleksi dokumen memperlihatkan bahwa pembobotan term dapat meningkatkan kinerja perolehan informasi yang didapatkan dengan term tanpa bobot menurut Salton, 1989 pada (Irmawati, 2017). Jadi, skema pembobotan istilah yang baik akan terlihat dalam hasil saat membedakan istilah yang diinginkan dan tidak diinginkan.

Perhitungan bobot term (*term weighting*) merupakan hal yang perlu diperhatikan di dalam melakukan pencarian informasi dari beberapa dokumen yang bersifat heterogen informasinya. Untuk mengetahui konteks dari suatu dokumen dapat dilakukan dengan mengidentifikasi dokumen tersebut berdasarkan term (kata) yang ada di dalam dokumen, dari frasa atau disebut sebagai hasil proses pengindeksan dokumen, oleh karena itu setiap kata perlu diberi suatu indikator yang disebut sebagai *term weight*

Pembobotan istilah dalam index rangking pada umumnya dibagi menjadi tiga katagori sebagai berikut :

1. Frekuensi istilah (*term frequency*) $tf_{i,j}$: banyaknya kemunculan istilah t_i dalam dokumen d_j. Istilah-istilah yang mempunyai frekuensi lebih tinggi dalam sebuah dokumen dipertimbangkan sebagai descriptor yang lebih baik dari isi sebuah dokumen.
2. Frekuensi dokumen (*document frequency*) df_i : banyaknya dokumen dalam koleksi dimana istilah t_i muncul di dalamnya. Sebuah istilah yang relevan sering muncul beberapa kali dalam sebuah dokumen. Di sisi lain, istilah yang tidak relevan sering muncul secara homogen dalam semua dokumen.
   Dari contoh koleksi dokumen, akan dilakukan perhitungan frekuensi dokumen dari term "Informasi", term "Sistem", "Akademik" dan "Perancangan".
3. Frekuensi koleksi (*collection frequency*) cf_i : banyaknya kemunculan istilah t_i dalam koleksi. Dari contoh koleksi dokumen, akan dihitung frekuensi koleksi (cf_i) frekuensi koleksi dari term term "Informasi", "Sistem", "Akademik", dan "Perancangan.

Kesamaan antara dokumen d_i dengan dokumen *dj* dapat diukur dengan fungsi similaritas (mengukur kesamaan) atau fungsi jarak (mengukuran ketidaksamaan). Salah satu fungsi similaritas (pengukuran kesamaan) antar dokumen adalah *eucledian distance*. *Eucledian distance* adalah perhitungan jarak dari 2 buah dokumen, metode perhitungan jarak kedekatan antar dokumen dihitung dengan nilai jarak dari 2 buah variable. Metode *ecluedian distance* lebih mudah dan efisiensi dalam proses. *Euclidean Distance* (*euclid*) Jarak *Euclid* adalah akar dari jumlah kuadrat perbedaan nilai untuk tiap variabel. Ukuran jarak antar objek ke-*i* dengan objek ke-*j* disimbolkan dengan d_{ij} dan $k = 1, 2, 3, \ldots,$. Nilai d_{ij} dapat dihitung dengan persamaan:

$$dij = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (2)$$

Keterangan:
d_{ij} = jarak antar objek ke-*i* dan objek ke-*j*
p = jumlah variabel cluster
x_{ik} = nilai atau data dari objek ke-i pada variabel ke-*k*
x_{jk} = nilai atau data dari objek ke-j pada variabel ke-*k*

c. Hierarchical Clustering

Penelitian yang terkait dengan clustering document sudah banyak dilakukan, pada dasarnya yang dimaksud dengan klastering dokumen yaitu proses mengelompokan (klaster) dokumen dengan mempertimbangkan kemiripan antar dokumen (document matching), yaitu antara dokumen yang satu dangan lainnya yang ada di dalam satu klaster menurut Ellis, 1996 dalam penelitian (Februariyanti & Santoso, 2017). Menurut (Zhang et al., 2001) Klastering dokumen bertujuan untuk memisahkan dokumen dari dokumen relevan dengan dokumen yang tidak relevan. Disebut juga bahwa dokumen yang relevan dengan kata kunci (*query*) lebih memiliki kecenderungan mirip (*matching*) antara satu dokumen dangan lainnya dibandingkan dengan dokumen yang tidak relevan, dengan demikian maka dokumen yang memiliki kemiripan dapat dikelompokan dalam sebuah klaster.

Pembentukan dari suatu klaster memiliki metode yang biasanya dikategorikan berdasarkan tipe dari klaster yang dihasilkan. Metode pembentukan klaster secara umum dibagi menjadi 2 (dua) yaitu klastering non hirarkis (*Non Hierarchical Clustering*) dan metode klastering hirarkis (*Hierarchical Cluastering*).

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

Metode klastering *Non Hierarchical Clustering* seringkali disebut sebgai metode partisi, yaitu metode klastering dengan cara membagi beberapa data yang meliputi n-obyek kedalam k-klaster dimana k < n dan obyek-obyek dalam klaster tidak saling *overlap* serta nilai k yang merupakan klaster sebelumnya telah ditentukan. Salah satu contoh metode non-hirarkis dalam klastering adalah metode k-means. K-means bertujuan untuk mengelompokan suatu obyek sedemikian rupa sehingga jarak dokumen yang satu dengan dokumen yang lain atau jarak masing-masing obyek ke pusat klaster dalam suatu klaster tertentu adalah minimum.

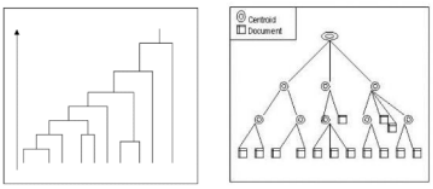
Pembentukan klaster dokumen menurut Salton, 1998 seperti pada penelitian (Februariyanti & Winarko, 2010) dalam *Information Retrieval* (Sistem Temu Kembali Informasi) yang dimaksud dengan metode klastering non hirarkis adalah sebagai berikut :

1. Melakukan pembandingan ciri identifikasi (*identifier)* dasri satu dokumen dengan dokumen ainnya dalam suatu koleksi dokumen serta mengkelompokkan dokumen-dokumen yang teridentifikasi memiliki ciri-ciri yang sama (serupa dalam sebuah klaster.
2. Disetiap klaster dokumen yang didapatkan, akan dipilih suatu *centroid* yaitu merupakan salah satu dokumen yang akan mewakili dokumen lainnya yang berada dalam klaster yang sama. *Centroid* adalah dokumen yang akan mewakili klaster dokumen yang merupakan record yang memiliki karakteristik atau ciri-ciri dokumen didalam sebuah klaster.
3. Penelusuran suatu dokumen dilakukan dalam dua proses yaitu :
   a. Dengan cara melalukan perbandingan query pada masing-masing klaster dokumen dengan *centroid*
   b. Melakukan pencocokan *query* pada dokumen masing-masing dokumen yang ada di dalam klaster yang paling sesuai dengan centroidnya.

Sedangkan untuk membentuk klaster dokumen menggunakan metode hirarkis dalam Sistem Temu Kembali Informasi adalah dengan cara :

1. Langkah pertama adalah mengidentifikasi dua buah dokumen, yaitu dokumen yang memiliki kemiripan paling tinggi yang selanjutnya digabungkan menjadi satu klaster
2. Lengkah kedua adalah melakukan identifikasi dan menggabungkan dua dokumen yang memiliki kemiripan paling tinggi berikutnya, selanjutnya akan dijadikan sebuah klaster berikutnya sehingga semua dokumen yang ada telah tergabung dan terbentuk sebuah klaster
3. Proses penelusuran dokumen dilakukan dengan cara melakukan pencocokan query dokumen dengan centroid yang dipilih. Centroid adalah dokumen yang dipilih ayang akan dijadikan sebagai dokumen parent dari masing-masing klaster tiap dokumen. Dokumen hasil dari query merupakan dokumen yang berada dalam satu klaster dengan centroid yang dipilih.

Secara grafik hasil algoritma *Hierarchical Clustering* secara grafik digambarkan sebagai suatu *tree*, yang sering disebut juga dengan dendogram. Secara grafik *tree* ini akan menggambarkan proses penggabungan dari klaster-klaster yang ada, sehingga akan menghasilkan klaster dengan level yang lebih tinggi dari klaster sebelumnya. Cabang-cabang dari *tree* akan menyajikan klaster. Selanjutnya cabang-cabang tergabung dalam suatu dode dengan posisi node tersebut bergabung sepanjang jarak yang mirip yang dihitung bersadarkan similaritas dokumen yang akan menggambarkan klaster yang terjadi. Gambar 1 berikut diperlihatkan struktur dari dendogram dan diagram *tree* untuk klaster hirarkhis.



Gambar 1 Dendogram dan Struktur Pohon dari *Hierarchical Clustering*

Untuk menentukan kesamaan atau kemiripan dari suatu dokumen diukur dengan cara menghitung jarak antar dokumen. Dokumen akan dikatakan memiliki kemiripan paling tinggi jika dua dokumen tersebut memiliki jarak paling kecil, maka akan dikelompokan menjadi satu klaster yang sama. Sebaliknya dokumen akan dikatakan memiliki kesamaan paling rendah jika dua dokumen tersebut memiliki jarak paling besar, selanjutnya akan dimasukkan dalam klaster yang berbeda

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

d. Single-Linkage Hierarchical Clustering

Ada tiga metode klaster hirarkhis, yaitu : *singel linkage*, metode klastring yang akan memberikan hasil jika klaster dihasilkan dari penggabungan berdasarkan jarak terdekat antar anggota yang satu dengan lainnya, *complete linkage*, yaitu metode klaster dimana klaster yang terbentuk merupakan penggabungan antara anggota-anggota yang memiliki jarak paling jauh dan *average linkage* adalah metode klaster yang akan menggabungkan jarak rata-rata diantara anggota-anggotanya pada himpunan klaster yang akan terbentuk.

Metode *single linkage* untuk pembentukkan klaster dokumen oleh Salton, 1989 dalam (Irmawati, 2017) input untuk algoritma *single linkage* adalah jarak atau sering disebut sebagai kesamaan (similaritas) antara pasangan-pasangan dari objek-objek. Klaster akan terbentuk dari entitas tunggal dengan cara menggabungkan entitas yang memiliki jarak paling rendah (pendek) yang disebut sebagai entitas dengan memiliki similaritas (kemiripan) yang paling tinggi. Pertama-tama kita harus menemukan jarak terpendek dalam $D = \{d_{ik}\}$ dan menggabungkannya dalam objek-objek yang bersesuaian misalnya, U dan V , untuk mendapatkan klaster (UV). Untuk langkah (3) dari algoritma di atas jarak-jarak antara (UV) dan klaster W yang lain dihitung dengan cara :

$$d_{(uv)w} = min\{ d_{uw}, d_{vw} \} \quad (1)$$

Nilai dari d_{UW} dan d_{VW} *merupakan* jarak paling pendek diantara klaster-klaster U dan W serta klaster-klaster dari V dan W.

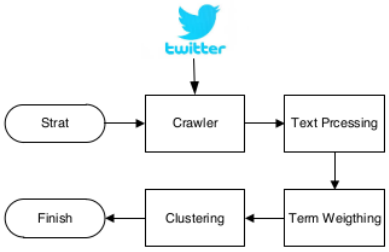
2. METODOLOGI PENELITIAN

a. Obyek Penelitian

Obyek penelitian dari penelitian ini dokumen informasi tweet yang ada pada akun twitter @unisbank dari tanggal 11 Agustus 2018 sampai 17 Juli 2019.

b. Tahapan Penelitian

Tahapan penelitian yang dilakukan dapat dilihat pada gambar 3 di bawah ini :



Gambar 2. Tahapan Penelitian

Tahapan penelitian dapat dijelaskan sebagai berikut:

1. **Proses *Crawler***
   Tahap awal penelitian adalah mengumpulkan data penelitian dengan cara proses crawler dari aplikasi crawling hasil penelitian terdahulu (Zuliarso & Mustofa, 2009) (Februariyanti et al., 2010) untuk mengambil data dari akun twitter secara otomatis
2. ***Text Processing***
   Selanjutnya setelah data didapatkan akan dilakukan text prosesing. Yaitu proses pembersihan data. *Text processing* dimulai dengan proses pembentukan kata mencadi *lowercase*. Selanjutnya dilakukan proses menghilangkan tanda baca yang ada pada dokumen. Kemudian dilanjutkan dengan prosen *tokenizing* dan *stopword removal*.
3. **Proses *Term Weigthing***
   Dari dokumen hasil dari *text processing* dilanjutkan dengan proses term weigthing, yaitu proses pembobotan dokumen dengan menggunakan TF-IDF dilanjutkan dengan similaritas dokumen dengan menghitung jarak antar dokumen dengan menggunakan algoritma *eucledian distance*.
4. ***Clustering Document***
   Dari hasil perhitungan similartas dokumen dilakukan proses klastering dokumen (*clustering document*) dengan menggunakan algoritma *Hieararchical Agglomerative*, serta ditampilkan hasil clastering dalam diagram dendogram.

3. HASIL DAN PEMBAHASAN

a. Data Set

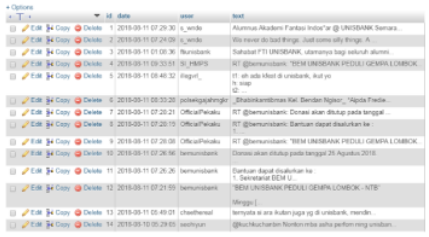
Data penelitian ini diambil dari dokumen informasi tweet yang ada pada akun twitter @unisbank dari tanggal 11 Agustus 2018 jam 07:29 wib sampai 17 Juli 2019 jam 06.12 wib dengan menghasilkan data tweet sejumlah 82 tweet. Dari proses crawler ternyata tidak banyak yang aktif di twitter @unisbank. Dari penelitian

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

yang telah dilakukan menghasilkan prototipe aplikasi *Crawler*. Proses pengambilan dokumen tweet dilakukan dengan cara mendapatkan dokumen di twitter yaitu berupa *data analytics* dengan *crawling* data twitter. Hasil proses crawler di twitter diperlihatkan pada gambar 3 dibawah ini:

Gambar 3 Hasil Crowler

Penyimpanan data merupakan tahap selanjutnya setelah melakukan proses penarikan data atau crawling data Twitter. Setelah proses crawler selanjutnya adalah proses penyimpana data hasil crawler ke dalam Database (basis data). Basis Data yang digunakan pada proses penyimpanan ini ialah twitterDB. Proses crowler pada penelitian ini dilakukan dengan cara update data secara otomatis. Di twitter disediakan fasilitas pengambilan data dengan kurun waktu 2 (dua minggu) yaitu secara otomatis dalam kurun waktu 2 (dua) minggu twitter akan melakukan update data. Untuk itu agar data dapat dianalisis dalam penelitian ini telah dilakukan pengambilan data secara otomatis setiap 2 (dua) minggu dan data akan tersimpan di database twitter dalam tabel tweet yang telah dibuat dapat dilihat pada gambar 4.



Gambar 4. Database Hasil Crawler

Dari data hasil crawler yang telah tersimpan di database twitter selanjutnya dilakukan proses filtering-folding. Yaitu proses proses pembentukan dokumen teks menjadi lowercase. Hasil dari proses filtering-folding dapat dilihat pada gambar 5

Gambar 5. Hasil Proses Filtering-Folding

Setelah dilakukan proses pembentukan dokumen teks menjadi lowercase semua dilanjutkan dengan proses Filtering-Punctuation. Yaitu proses pembersihan dokumen teks dari tanda baca yang ada. Hasil proses Filtering-Punctuation penelitian ini dapat dilihat pada gambar 6 di bawah ini :

Gambar 6. Hasil Proses *Filtering-Punctuation*

Dari proses Filtering-Punctuation dilanjutkan dengan proses tokenizing, yaitu proses pemotogan kata (term). Proses tokenizing ini dilakukan dengan cara memotong term berdasarkan spasi, jika term tersebut dipisahkan oleh spasi maka akan dianggap sebuah term sendiri. Hasil prosesn tokenizing pada masing-masing dokumen tweet diperlihatkan pada gambar 7 di bawah ini :

| Koneksi | Crawling | Tables | Filtering-Folding | Filtering-Punctuation

Koneksi Berhasil

No	ID	Date	User	Text
1	1	2018-08-11 07:29:30	s_wnde	akademi
2	1	2018-08-11 07:29:30	s_wnde	fantasi
3	1	2018-08-11 07:29:30	s_wnde	indosiar
4	1	2018-08-11 07:29:30	s_wnde	unisbank
5	1	2018-08-11 07:29:30	s_wnde	semarang
6	1	2018-08-11 07:29:30	s_wnde	alumnus
7	4	2018-08-11 09:33:51	SI_HMPS	minggu
8	2	2018-08-11 07:24:09	s_wnde	mean
9	2	2018-08-11 07:24:09	s_wnde	unisbank
10	2	2018-08-11 07:24:09	s_wnde	semarang
11	2	2018-08-11 07:24:09	s_wnde	lot
12	2	2018-08-11 07:24:09	s_wnde	never
13	2	2018-08-11 07:24:09	s_wnde	do

Gamber 7. Dokumen Hasil Proses Tokenizing

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

Selanjutnya akan dilakukan dengan proses stopword removal. Stopword Removal. Yaitu proses menghilangkan kata-kata yang tidak bermakna atau tidak ada arti dari kata tersebut. Proses dilakukan dengan cara mencocokan dokumen hasil twitter yang telah dilakukan proses tokenizing dengan mencocokan teks yang telah tersimpan di table stopword sebagai term stopword. Jika teks yang dimakasudkan sama dengan kata dalam table stopword makan akan dihapus. Data hasil dari proses stopword removal seperti terlihat pada gambar 9

| Koneksi | Crawling | Tables | Filtering-Folding | Filtering-Pu

Koneksi Berhasil

No	ID	Date	User	Text
1	1	2018-08-11 07:29:30	s_wnde	akademi
2	1	2018-08-11 07:29:30	s_wnde	fantasi
3	1	2018-08-11 07:29:30	s_wnde	indosiar
4	1	2018-08-11 07:29:30	s_wnde	unisbank
5	1	2018-08-11 07:29:30	s_wnde	semarang
6	1	2018-08-11 07:29:30	s_wnde	alumnus
7	4	2018-08-11 09:33:51	SI_HMPS	minggu
8	2	2018-08-11 07:24:09	s_wnde	mean
9	2	2018-08-11 07:24:09	s_wnde	unisbank
10	2	2018-08-11 07:24:09	s_wnde	semarang
11	2	2018-08-11 07:24:09	s_wnde	lot
12	2	2018-08-11 07:24:09	s_wnde	never

Gambar 8. Hasil Proses Stopword Removal

Dokumen teks hasil dari dari proses stopword removal dan disimpan dalam tabel selanjutnya akan dilakukan analisis sehingga dihasilkan dokumen yang bersih dari noise. Dokumen yang sudah bersih dari noise selanjutnya dapat digunakan sebagai dokumen untuk uji coba dalam penelitian. Dokumen yang telah dihasilkan dan digunakan sebagai informasi akan lebih mudah jika divisualisasikan dalam bentuk peta. Selanjutnya proses penyimpanan harus dilakukan secara langsug atau sering disebut dengan direct storing. Proses tersebut dilakukan dikarenakan data yang diambil dari twitter merupakan data realtime, sehingga perlu adanya sebuah database yang dapat untuk menyimpan dokumen hasil penarikan secara langsung (direct storing)

b. Perhitungan TF-IDF

Metode TF-IDF adalah proses pemberian bobot dari suatu term (kata) dalam suatu dokumen, metode TF-IDF merupakan metode yang menggabungkan dua konsep untuk proses pembobotan dokumen. Perhitungan dilakukan dengan cara menghitung frekuensi kemunculan suatu kata (term) dalam suatu dokumen serta menghitung *Inverce Document Frequency* (IDF) yang mengandung kata yang dimaksudkan. Frekuensi kemunculan term dalam suatu dokumen akan memberikan informasi seberapa penting kata tersebut dalam dokumen yang dimaksud. Sedangkan frekuensi dari dokumen yang mengandung kata yang dimaksud akan menunjukkan seberapa umum kata tersebut dalam dokumen yang dimaksud. Dengan demikian maka bobot keterkaitan antara suatu kata dan suatu dokumen akan bernilai tinggi jika frekuensi kata tersebut juga bernilai tinggi dalam dokumen tersebut, serta frekuensi semua dokumen yang mengangandung kata yang dimaksud bernilai rendah di kumpulan dokumen yang dimasudkan. Dengan perhitungan TF-IDF akan menghasilkan perhitungan yang dapat dilihat pada tabel 1 di bawah :

Tabel 1. Hasil Perhitungan TF-IDF

Id	No	Term	TF	DF	IDF
1	1	f	1	2	1.6127838567197355
2	1	akademi	1	1	1.9138138523837167
3	1	semarang	1	16	0.7096938697277919
4	1	fantasi	1	1	1.9138138523837167
5	1	alumnus	1	1	1.9138138523837167
6	1	t	1	46	0.2510560205432544
7	1	https	1	46	0.2510560205432544
8	1	ar	1	1	1.9138138523837167
9	1	unisbank	1	67	0.08773904967759397
10	1	indos	1	1	1.9138138523837167

c. Hitung Jarak Euclidean Distance

Kesamaan antara dokumen Di dengan dokumen Dj dapat diukur dengan fungsi similaritas (mengukur kesamaan) atau fungsi jarak (mengukuran ketidaksamaan). Salah satu fungsi similaritas (pengukuran kesamaan) antar dokumen adalah eucledian distance. Eucledian distance adalah perhitungan jarak dari 2 buah dokumen, metode perhitungan jarak kedekatan antar dokumen dihitung dengan nilai jarak dari 2 buah variable. Metode ecluedian distance lebih mudah dan efisiensi dalam proses. Hasil perhitungan jarak antar 2 (dua) dokumen dapat dilihat pada tabel 2 dan table 3 :

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

Tabel 2 Hitung Jarak Dokumen 1

A	B	Ecludien
doc2	doc1	47
doc3	doc1	46
doc4	doc1	46
doc5	doc1	43
doc6	doc1	46
doc7	doc1	34
doc8	doc1	31
doc9	doc1	46

Tabel 3 Hitung Jarak Dokumen 2

A	B	Ecludien
doc1	doc2	26
doc3	doc2	46
doc4	doc2	46
doc5	doc2	43
doc6	doc2	46
doc7	doc2	34
doc8	doc2	31

Pada tabel 2 diperlihatkan perhitungan jarak dokumen 1 dengan dokumen yang lain, yaitu dokumen 2, dokumen 3 dan seterusnya. Sedangkan tabel 3 diperlihatkan hasil perhitungan jarak dokumen 2 dengan dokumen lainnya yaitu dokumen 1, dokumen 3 dan seterusnya. Dari hasil perhitungan jarak pada tabel 2 dan 3 dapat dibuat matrik hasil perhitungan jarak antar dokumen seperti terlihat pada tabel 4 dibawah ini :

Tabel 4 Matrik Jarak Antar Dokumen

	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6
1		26	28	34	28	28
2	47		49	55	49	49
3	46	46		52	46	46
4	46	46	46		46	46
5	43	43	43	49		43
6	46	46	46	52	46	

d. Hierarchical Clustering

Hierarchical Clustering merupakan metode dalam menganalisis klaster dengan cara membentuk suatu hirarki klaster data (dokumen). Cara pembentukkan kelompok terdiri dari dua metode yaitu dengan metode Agglomerative (Bottom-Up) dan Devisive (Top-Down).

Tahapan di dalam algoritma Agglomerative Hierarchical Clustering adalah sebagai berikut :

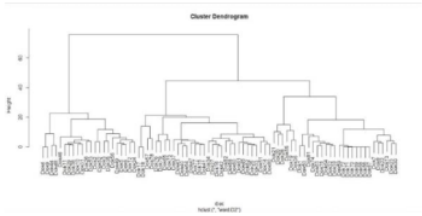
1. Pertama akan **di**hitung Matrik Jarak antar dokumen.
2. Langkah kedua dilakukan penggabungkan dua klaster paling dekat dengan berdasar pada parameter kedekatan yang telah ditentukan.
3. Langkah ketiga perbarui Matrik Jarak antar dokumen yang akan digunakan untuk merepresentasikan jarak antara klaster yang baru dan klaster yang masih ada.
4. Selanjutnya ulangi proses ke-2 dan proses ke-3 sehingga hanya akan didaptkan satu klaster yang tersisa.

Dari hasil klastering menggunakan metode hierarchical agglomerative dihasilkan klaster paling banyak dengan jarak kedekatan (simmilaritas) dokumen sebesar 28. Tabel hasil klaster terbanyak dapat diperlihatkan pada tabel 5. di bawah ini :

Tabel 5 Hasil Klaster Terbanyak

A	B	Ecludien1
doc59	doc1	28
doc66	doc1	28
doc67	doc1	28
doc68	doc1	28
doc69	doc1	28

Dari matrik hitung jarak dokumen yang dihasilkan dapat dilakukan klastering menggunakan Hierarchical Agglomerative Clustering sehingga didapkan hasil dendogram seperti terlihat pada gambar 10 Klaster terbanyak dihasilkan dari dokumen yang menggunakan kata-kata Unisbank Semarang.



Gambar 10. Dendogram Hierarchical Agglomerative Clustering.

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 13 No. 1, Mei 2021
eISSN : 2580-3042
pISSN : 1979-0694

4. KESIMPULAN

Dari hasil penelitian dan uji coba yang telah dilakukan, maka dapat diambil kesimpulan bahwa kedekatan atau kemiripan (similaritas) dengan mengukur jarak menggunakan algoritma *eucledian distance* dihasilkan jarak terdekat antar dokumen yang satu dengan yang lainnya adalah 1 dan jarak terjauh 61. Selanjutnya dengan menggunakan *Hierarchical Clustering* didapatkan hasil klastering paling banyak tweet dengan menggunakan kata "Unisbank Semarang" dengan sejumlah dokumen dan Cluster yang terbentuk, belum mencerminkan topik diskusi yang spesifik dalam memanfaatkan twitter di akun @unisbank.

5. REFERENSI

Coletta, L. F. S., De Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Combining classification and clustering for tweet sentiment analysis. *Proceedings - 2014 Brazilian Conference on Intelligent Systems, BRACIS 2014*, *July*, 210–215. https://doi.org/10.1109/BRACIS.2014.46

Februariyanti, H., & Santoso, D. B. (2017). Hierarchical Agglomerative Clustering Untuk Pengelompokan Skripsi Mahasiswa. *Seminar Nasional Teknologi Informasi Dan Aplikasi Komputer*, 33–40. https://doi.org/10.1016/0031-3203(79)90049-9

Februariyanti, H., & Winarko, E. (2010). Klastering Dokumen Menggunakan. *Seminar Nasional STIKOM Surabaya*. https://media.neliti.com/media/publications/220918-klastering-dokumen-menggunakan-hierarchi.pdf

Februariyanti, H., Zuliarso, E., & Utomo, S. (2010). *Prototipe Mesin Pencari Dokumen Teks*. *XV*(2), 115–120. https://www.unisbank.ac.id/ojs/index.php/fti1/article/view/119

Irmawati, I. (2017). Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model. *Jurnal Ilmiah FIFO*, *9*(1), 74. https://doi.org/10.22441/fifo.v9i1.1444

Nugroho, E. (2011). *Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Menggunakan Algoritma Rabin-Karp* (Issue Januari). http://blog.ub.ac.id/ecoorner/files/2011/03/Bab12345.pdf

Prabawati, R. L., & Dawud. (2019). Karakteristik Argumentasi Dalam Opini di Media Online. *Basindo*, *3*(2), 224–236. http://journal2.um.ac.id/index.php/basindo/article/view/11586/5015

Susandi, D., & Sholahudin, U. (2016). Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani , KNN dan Fungsi Similarity Cosine untuk Pembobotan IDF dan WIDF pada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia. *Jurnal ProTekInfo*, *3*(1), 22–29. http://download.garuda.ristekdikti.go.id/article.php?article=822075&val=13339

Tan, A. (2011). Text Mining : The state of the art and the challenges Concept-based. *Proceedings of the PAKDD 1999 Workshop On*, *March*, 65–70. http://www.mendeley.com/research/text-mining-state-art-challenges-3/

Zhang, J., Gao, J., Zhou, M., & Wang, J. (2001). Improving the Effectiveness of Information Retrieval with Clustering and Fusion. *Computational Linguistics and Chinese Language Processing*, *6*(1), 109–125. http://www.aclclp.org.tw/clclp/v6n1/v6n1a5.pdf

Zuliarso, E., & Mustofa, K. (2009). Crawling Web berdasarkan Ontology. *Jurnal Teknologi Informasi DINAMIK*, *XIV*(2), 105–112. https://repository.ugm.ac.id/32972/1/97-283-1-SM_(1).pdf

ANALISIS KECENDERUNGAN INFORMASI

ORIGINALITY REPORT

18%	15%	2%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ es.scribd.com

Internet Source

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		