ANALISIS KLASTER PADA DOKUMEN TEKS OPINI PENGGUNA TWITTER TERHADAP KASUS MIRAS OPLOSAN MENGGUNAKAN METODE K-MEANS

[1]Jaka Aulia Pratama, [2]Neneng Sunengsih, [3]Maman Suherman
[1]Bidang Statistika dan Persandian, DISKOMINFO Kabupaten Bandung
[2]Program Studi Statistika, FMIPA, Universitas Padjadjaran Bandung
[3]Fakultas Ilmu Sosial dan Politik, Universitas Nurtanio Bandung
Email : jakajek@gmail.com

ABSTRAK

Teknologi komunikasi dan informasi merupakan sektor yang paling pesat berkembang di era digital saat ini. Hal tersebut tidak lepas dari kebutuhan mendasar manusia sebagai makhluk sosial, dimana akses terhadap informasi dan keragaman bentuk dalam berkomunikasi menjadi lahan basah bagi para penyedia layanan, salah satunya *Twitter*. Layanan jejaring sosial *Twitter* menjadi wadah dalam menyampaikan berbagai macam opini, termasuk kasus miras oplosan yang *viral* disampaikan para penggunanya selama bulan April 2018. Penelitian ini bertujuan untuk menganalisi opini pengguna *Twitter* terhadap kasus mira oplosan di bulan April 2018 tersebut menggunakan metode *K-means*. Hasil dari penelitian ini menunjukan klaster paling optimum terbentuk sebanyak tiga klaster berdasarkan nilai *dunn index* sebesar 0.8312. Dari ketiga klaster tersebut, dapat diasumsikan opini pengguna *Twitter* dari tanggal 1 April 2018 hingga 23 April 2018 terhadap kasus miras oplosanmasih terpusat pada sosok pengedar miras oplosan, pihak berwenang, dan korban.

Kata kunci : *Text Mining*, Klaster, *K-means*,*Twitter*

PENDAHULUAN

Perkembangan teknologi komunikasi dan informasi di dunia digital yang begitu pesat dirasakan saat ini, merupakan indikasi dari kebutuhan manusia sebagai makhluk sosial yang menginginkan cara efektif juga efisien dalam berkomunikasi dan mendapatkan informasi. Salah satu wadah komunikasi dan informasi dalam dunia digital antara lain adalah jejaring sosial.Berbagai penyedia layanan jejaring sosial bermunculan seiring berkembangnya cara manusia berkomunikasi. Oleh karenanya, setiap penyedia layanan jejaring sosial berlomba untuk menawarkan fitur-fitur yang berbeda antar satu dan lainnya, sehingga masing-masing penyedia layanan jejaring sosial memiliki fitur unik yang menjadi pilihan bagi penggunanya. Salah satu layanan jejaring sosial yang memiliki fitur unik dalam layanan komunikasi dan informasi adalah layanan jejaring sosial *Twitter*. *Twitter* merupakan suatu wadah berkomunikasi dan berbagi informasi, dimana bentuk komunikasi dan informasi dapat disampaikan dalam sebuah *tweet*.

Banyaknya karakter dalam sebuah *tweet* dibatasi sebanyak seratus empat puluh karakter, oleh karenanya pengguna layanan jejaring sosial *Twitter* dituntut untuk menggunakan kata-kata yang singkat, padat dan jelas dalam berkomunikasi dan berbagi informasi

dengan sesama pengguna. Informasi-informasi yang biasa dibagikan di layanan jejaring sosial *Twitter* sangat beragam bentuknya. Salah satu bentuk informasi antara lain adalah opini terhadap suatu fenomena. *Twitter* menyediakan fitur *hashtag* (#), dimana fitur tersebut berfungsi untuk menjadikan suatu kata sebagai kata kunci dalam suatu opini dengan topik yang sama. Salah satu opini yang sedang ramai disampaikan selama bulan April 2018 melalui *tweet* dan *hashtag* antara lain mengenai fenomena miras oplosan di Indonesia. Banyaknya korban dan mulai terungkapnya bisnis miras oplosan di Indonesia, mengakibatkan layanan jejaring sosial *Twitter* menjadi wadah bagi para penggunanya untuk menyampaikan beragam opini dengan topik tersebut. Beragamnya opini pada topik miras oplosan di layanan jejaring sosial *Twitter* selama April 2018,menjadi suatu kumpulan data berupa teks yang dapat dianalisa lebih dalam melalui pendekatan statistika.

Dari kumpulan data teks berupa opini tersebut, salah satu informasi yang bisa didapatkan adalah kata apa yang sering muncul pada kumpulan opini tersebut, dan kata-kata tersebut dapat terbagi kedalam beberapa klaster. Oleh karenanya, pada penelitian ini peneliti bermaksud untuk mengaplikasikan *Text Mining* pada kumpulan data teks berupa opini pengguna layanan jejaring sosial *Twitter* terhadap miras oplosan selama bulan April 2018. Sedangkan tujuan dalam penelitian ini adalah mengetahui kata-kata yang sering muncul, dan mendapatkan klaster-klaster kata yang terbentuk berdasarkan kumpulan opini pengguna *Twitter* terhadap miras oplosan selama bulan April 2018.

Analisis klaster atau *clustering* merupakan suatu proses mengelompokan data ke dalam sebuah kelas atau *cluster*, dimana objek yang berada di dalam klaster memiliki tingkat kemiripan yang tinggi satu sama lainnya tetapi memiliki tingkat ketidakmiripan yang tinggi dengan objek di klaster lain. Metode *clustering* dibagi menjadi dua kelompok, yaitu metode hierarki dan metode non-hierarki [3]. Pada metode non-hierarkijumlah klaster *(k)* sudah ditentukan diawal pengelompokan. Metode non-hierarkidapat diaplikasikan untuk set data yang sangat lebih besar jika dibandingkan dengan metode hierarki. Menurut [4], teknik *clustering* non-hierarkidirancang untuk mengelompokan item bukan variabel. Salah satu algoritma pada metode *clutering non-hierarki* yaitu *K-means*.

K-means merupakan salah satu metode non-hierarkiyang paling sering digunakan [4]. Pada metode ini sebuah set data diklasifikasikan ke dalam beberapa klaster yang sudah ditentukan, dimana klaster tersebut diasumsuikan *fix.K-means* menggunakan *centroid* sebagai pusat klasternya, *centroid* tersebut biasanya berupa nilai rata-rata. *K-means* membagi objek kedalam *k* klaster, kemudian menempatkan objek kedalam klaster yang memiliki jarak *centroid* terdekat. Setelah itu dilakukan perhitungan kembali nilai *centroid* yang baru, dan dilakukan kembali penempatan objek kedalam klaster yang memiliki jarak terekat dengan nilai *centroid* yang baru. Proses tersebut terus dilakukan sampai tidak ada *centroid* yang berpindah. Pengukuran jarak dilakukan untuk menunjukan kedekatan antara dua buah objek. Pada penelitian ini, ukuran jarak yang akan digunakan adalah *Euclidean distance*, hal ini dikarenakan *Euclidean distance* merupakan *dissimilarity measure* yang paling umum digunakan untuk data yang berskala interval atau rasio [6].

METODE PENELITIAN

Text Mining

Text Mining bisa didefinisikan sebagai proses penggalian informasi di mana pengguna berinteraksi dengan kumpulan dokumen dari waktu ke waktu dengan menggunakan suatu alat analisis. *Text Mining* mencari informasi berarti dari sumber-sumber data melalui identifikasi dan eksplorasi pola tertentu, dalam kasus ini sumber data adalah kumpulan dokumen dengan pola yang ditemukan pada data teks yang tidak berstruktur. Praproses dari *Text Mining* sendiri berpusat pada identifikasi dan ekstraksi fitur representatif untuk dokumen *Natural Language* [2].

Proses *Text Mining* biasanya membutuhkan penyusunan teks masukan berdasarkan tata bahasa, yang diikuti dengan menggali pola dari data yang sudah terstruktur, evaluasi dan interpretasi hasil. Proses ini biasanya digunakan untuk pengklasifikasian, penggerombolan, analisis makna, pengambil kesimpulan dari dokumen dan pemodelan hubungan objek yang berupa kata [3]. Berikut merupakan tahapan dalam *Text Mining* menurut [5] :

1. *Information Retrieval*
   Yaitu tahapan untuk memperoleh dokumen yang sesuai dengan permintaan peneliti atau yang sesuai dengan permasalahan
2. *Naturtal Language Processing*
   Yaitu tahapan untuk mentransformasi kata-kata yang terdapat dalam dokumen yang telah diperoleh sebelumnya. Dimana dari dokeumen awalnya yang tidak terstruktur menjadi lebih terstruktur, sehingga dapat diperole informasi yang lebih akurat dan berguna.
3. *Information Extraction*
   Yaitu tahapan dimana informasi yang sudah diperoleh sebelumnya akan diekstrak sehingga peneliti akan lebih mudah memahami permasalahan yang diteliti melalui visualisasi yang ditampilkan
4. *Knowledge Discoovery*
   Pada tahapan ini pola dari suatu dokumen mulai teridentifikasi dan pengetahuan untuk mengatasi permasalahan telah didapat.

Dikarenakan data teks biasanya tidak terstruktur, terutama data teks yang diperoleh dari suatu layanan jejaring sosial, maka diperlukan tahapan persiapan data terlebuh dahulu. Hal ini bertujuan untuk mempermudah peneliti dalam menganalisis data teks tersebut. Berikut adalah tahapan persiapan data teks dalam penelitian ini:

1. *Tokenizing*
   Adalah proses penguraian *string* teks menjadi suatu *term* atau kata.
2. *Filtering*
   Adalah tahapan seleksi *tokens* yang dianggap tidak penting seperti *Stopwords. Stopwords* adalah kata-kata yang tidak mengandung makna penting seperti ‘dan’, ‘di’, ‘kamu’, ‘the’, ‘and’, dll.
3. *Stemming*
   Adalah suatu proses yang bertujuan untuk mengambil kata dasar dari kata yang mngandung imbuhan, baik itu imbuhan awalan, akhiran maupun awalan dan akhiran.

HASIL PENELITIAN

Pada bagian ini akan dijelaskan mengenai hasil analisa dari penelitian ini. *Tweets* dengan kata kunci “miras oplosan” yang diperoleh dari tanggal 1 April 2018 hingga 23 April 2018 adalah sebanyak 9954 *tweets.*Sebelum dianalisis menggunakan metode *K-means* klaster, *tweets* tersebut telah melewati serangkain prosedur persiapan data teks yaitu *tokenizing, filtering* dan *stemming*, sehingga *tweets* berbahasa Indonesia

tersebut sudah berupa kata dasar dan siap untuk dianalisa.

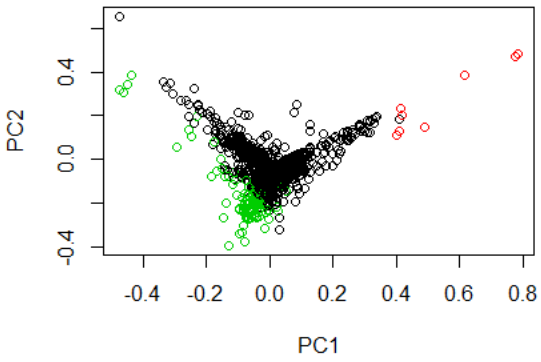
Berikut adalah Tabel 1 yang menunjukan hasil validasi klaster menggunakan *internal validationdunn index* untuk metode *K-means.*

Tabel 1. Validasi Klaster Menggunakan *Internal Validation Dunn Index*

Klaster	*Dunn Index*
2	0.8171
3	**0.8312**
4	0.5064
5	0.513
6	0.4811
7	0.4811
8	0.5074
9	0.5074
10	0.5188

Terlihat pada Tabel 1, klaster sebanyak tiga buah menghasilkan nilai *dunn index*yang paling tinggi diantara klaster-klaster lainnya yaitu sebesar 0.8312. Sejalan dengan hal tersebut, maka banyaknya klaster yang akan dibentuk pada penelitian ini adalah tiga klaster karena nilai *dunn index* tertinggi menunjukan bahwa klaster tersebut optimum. Dari tiga klaster yang terbentuk menggunakan metode *K-means* diperoleh *tweets* sebanyak 8318 untuk klaster satu, 420 *tweets* untuk klaster dua, dan 1216 *tweets* untuk klaster tiga.

Selanjutnya adalah Gambar 1 yang menunjukan *plot* komponen utama dari tiga klaster menggunakan metode *K-means* yang terbentuk.



Gambar 1. *Plot* Komponen Utama dari Tiga Klaster

Dari Gambar 1 tersebut, terlihat bahwa masih terdapat beberapa bulatan dengan warna berbeda yang masuk ke suatu kumpulan bulatan dengan warna dominan. Sehingga, pada tiga klaster tersebut masih terdapat beberapa *tweets* dan *terms* kata yang belum terpisah secara sempurna,walaupun nilai *dunn index* untuk tiga klaster menghasilkan nilai yang paling tinggi.Pada bagian selanjutnya, akan ditampilkan *terms* kata yang dominan dan *wordcloud*dari masing-masing klaster yang terbentuk.

Klaster Satu

Tabel 2. Sepuluh Kata Dominan Pada Klaster satu

Kata	Frekuensi
bos	1410
tewas	1391
orang	1326
minum	1102
korban	1011
polisi	982
edar	639
simbolon	596
maut	573
big	570

Gambar 2. *Wordcloud* Pada Klaster Satu

Tabel 2 menunjukan 10 kata yang paling sering muncul di klaster satu. Dari Tabel 2 tersebut, terlihat bahwa kata "bos" menjadi kata yang paling sering muncul pada*tweets*yang terdapat di klaster satu yaitu sebanyak 1410 kemunculan.Kemudian, Gambar 2 menunjukan visualisasi berupa *wordcloud* yang memperlihatkan kata-kata yang sering muncul di klaster satu. Terlihat pada Gambar 2, kata dengan ukuran yang semakin besar menunjukan bahwa kata tersebut adalah kata yang paling sering muncul di klaster satu.Pada bagian selanjutnya, akan ditampilkan *terms* kata yang dominan dan *wordcloud* dari klaster dua.

Klaster Dua

Tabel 3. Sepuluh Kata Dominan Pada Klaster Dua

Kata	Frekuensi
bos	85
minum	80
orang	80
mati	74
tewas	71
korban	67
narkoba	60
main	59
nama	59
tik	59

Gambar 3. *Wordcloud* Pada Klaster Dua

Tabel 3 menunjukan 10 kata yang paling sering muncul di klaster dua. Dari Tabel 3 tersebut, terlihat bahwa kata "bos" menjadi kata yang paling sering muncul pada *tweets* yang terdapat di klaster dua yaitu sebanyak 85 kemunculan. Kemudian, Gambar 3 menunjukan visualisasi berupa *wordcloud* yang memperlihatkan kata-kata yang sering muncul di klaster dua. Terlihat pada Gambar 3, kata dengan ukuran yang semakin besar menunjukan bahwa kata tersebut adalah kata yang paling sering muncul di klaster dua.Pada bagian selanjutnya, akan ditampilkan *terms* kata yang dominan dan *wordcloud*dari klaster tiga.

Klaster Tiga

Tabel 4. Sepuluh Kata Dominan Pada Klaster Tiga

Kata	Frekuensi
tewas	284
Bos	263
orang	260
korban	188
Big	160
minum	151
polisi	151
simbolon	151
syamsudin	146
edar	135

Gambar 4. *Wordcloud* Pada Klaster Tiga

Tabel 4 menunjukan 10 kata yang paling sering muncul di klaster tiga. Dari Tabel 4 tersebut, terlihat bahwa kata "tewas" menjadi kata yang paling sering muncul pada *tweets* yang terdapat di klaster tigayaitu sebanyak 284 kemunculan. Kemudian, Gambar 4 menunjukan visualisasi berupa *wordcloud* yang memperlihatkan kata-kata yang sering muncul di klaster tiga. Terlihat pada Gambar 4, kata dengan ukuran yang semakin besar menunjukan bahwa kata tersebut adalah kata yang paling sering muncul di klaster tiga.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan bahwa dapat disimpulkan data teks berupa *tweets* mengenai miras oplosan yang diambil pada tanggal 1 April 2018 hingga 23 April 2018 sebanyak 9954 *tweets*, dengan menggunakan metode *K-means* dapat dibagi menjadi tiga klaster dengan nilai *dunn index* sebesar 0.8312. *Plot* komponen utama dari ketiga klaster tersebut menunjukan adanya beberapa *tweets* yang tidak terpisah secara sempurna. Hal tersebut dapat terlihat juga pada kata-kata yang paling sering muncul di setiap klaster, yaitu adanya kata-kata yang sama muncul di masing-masing klaster. Pada data teks tersebut, masih ditemukan kata-kata yang tidak mengandung arti penting, hal ini dikarenakan kata-kata tersebut bukanlah bahasa Indonesia yang baku atau istilah-istilah terbarukan yang tidak tersaring oleh kamus *corpus* yang digunakan pada penelitian ini. Dari ketiga klaster yang terbentuk, opini pengguna layanan jejaring sosial *Twitter*terhadap miras oplosan dari tanggal 1 April 2019 hingga 23 April 2018, dapat diasumsikan masih terpusat pada sosok pengedar miras oplosan, pihak berwenang, dan korban.

Saran untuk penelitian selanjutnya adalah membandingkan hasil *K-means* klaster pada data teks berbahasa Indonesia dengan metode lainnya seperti *K-medoids* atau *Hierarchical*.Tahapan persiapan data seperti *cleaning* menjadi bagian yang sangat penting untuk menghindari kemunculan kata yang tidak mengandung arti penting, oleh karenanya diperlukan pengembangan lebih lanjut dalam kamus *corpus* yang digunakan pada penelitian ini seiring dengan berkembangnya istilah-istilah terbarukan atau bahasa "gaul" dan singkatan-singkatan dalam bahasa Indonesia.

DAFTAR PUSTAKA

[1] Ansari, Z. *et al*. 2011. *Quantitative Evaluation of Perfomance and Validity Indices for Clustering the Web Navigational Sessions*. World of Computer Science and Information Technology Journal. 5. 217-226.

[2] Feldman, R., & Sanger, J. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.

[3] J. Han, M. Kamber. 2006. *Data Mining Concepts and Techniques*, Morgan Kaufmann. America.

[4] Johnson, R A., Wichern, D W. 2002. *Applied Multivariate Statistical Analysis, the fifth edition*. Prentice Hall Inc. New Jersey.
[5] McDonald, Dr Diane. 2012. *The Value and Benefits of Text Mining.* JISC United Kindgom
[6] Timm, Neil H. 2002. *Applied Multivariat Analysis*: Springer.