JURNAL MATEMATIKA "MANTIK"
Edisi: Mei 2018. Vol. 04 No. 01
ISSN: 2527-3159 **E-ISSN: 2527-3167**

Text Mining dengan K-Means *Clustering* pada Tema LGBT dalam Arsip Tweet Masyarakat Kota Bandung

Eko Yulian[1]
Pusdiklat Badan Pusat Statistik Jakarta Selatan[1], okeyulian@gmail.com[1]

DOI:https://doi.org/10.15642/mantik.2018.4.1.53-58

Abstrak

Gerakan LGBT berkembang cepat melalui media sosial sehingga ide-ide LGBT dapat dengan leluasa dikemukakan. Tweeter merupakan salah satu media yang seringkali digunakan untuk tujuan tersebut. Komentar-komentar atau "cuitan" tentang LGBT di twitter tentu banyak jumlahnya. Banyaknya informasi yang ada di dunia maya membuat upaya-upaya pengembangan terhadap penggalian informasi dari basis data daring semakin pesat, salah satunya *text mining*. Salah satu teknik statistika yang bisa digunakan untuk memanfaatkan hasil dari *text mining* adalah *clustering*. *Clustering* yang digunakan pada penelitian ini adalah K-Means *clustering*. Penelitian ini menggunakan 5 *cluster* untuk mengelompokkan komentar-komentar di twitter yang berhubungan dengan LGBT di kota Bandung. Dari lima *cluster* yang dibentuk pada proses K-means diperoleh bahwa kecenderungan cuitan pengguna *Tweeter* kota bandung terkait LGBT secara umum masih berhubungan dangan perspektif religi yang ditandai dengan kemunculan kata agama yang sangat sering.

Kata kunci : K-Means Clustering, LGBT, Text Mining

Abstract

The movement of LGBT is growing rapidly through social media so that LGBT ideas can be freely expressed. The tweeter is one of the media that is often used for that purpose. Comments or "cuitan" about LGBT on twitter certainly many in number. The amount of information available in cyberspace makes development efforts to extract information from online databases rapidly, one of which is text mining. One of the statistical techniques that can be used to utilize the results of text mining is clustering. Clustering used in this study is K-Means clustering. This study uses 5 clusters to group comments on The twitter associated with LGBT in the city of Bandung. Of the five clusters formed in the K-means process, it is found that the tendency of Tuet Tweeter users of LGBT related bands in general, is still related to the religious perspective which is marked by the emergence of the word religion very often.

Keyword : K-Means Clustering, LGBT, Text Mining

ISSN: 2527-3159 E-ISSN: 2527-3167

1. Pendahuluan

LGBT atau GLBT adalah akronim dari "lesbian, gay, biseksual, dan transgender". Istilah ini mulai sering digunakan tahun 1990-an sebagai pengganti frasa yang lebih dulu populer, "komunitas gay", karena lebih mewakili kelompok- kelompok yang telah disebutkan. Kaum LGBT merupakan salah satu kelompok orang yang memiliki orientasi seksual sebagai homoseksual atau penyuka sesama jenis yang terjadi pada kaum pria [1]. Lebih lanjut kaum homoseksual didominasi oleh kaum laki-laki karena beberapa faktor seperti kelainan genetika dan faktor sosial seperti lingkungan yang memang mendukung untuk terbentuknya kaum tersebut atau karena terjadinya trauma dalam hubungan seksualitasnya [1].

Pemberitaan LGBT di Indonesia mulai marak di Indonesia setelah Mahkamah Agung Amerika Serikat melegalkan pernikahan sesama jenis pada 26 Juni 2015. Sejak saat itu, muncullah pemberitaan di media massa pada akhir tahun 2015 bahwa telah terjadi pernikahan sesama jenis di Indonesia. LGBT merupakan salah satu isu penting yang saat ini sedang marak diberitakan oleh beberapa media bahkan menjadi bahan diskusi oleh beberapa pakar di Indonesia. Contohnya majalah Gatra edisi 4-10 Februari 2016 [2] yang memberitakan isu "Arus LGBT Masuk Kampus di Indonesia",

Bandung sebagai salah satu kota pelajar sekaligus metropolitan yang menjadi tujuan belajar bagi para mahasiswa dan wisata unggulan di Indonesia tidak lepas sasaran perubahan pola sosial budaya yang begitu cepat termasuk kemunculan komunitas-komunitas LGBT. Berdasarkan catatan Badan Kesatuan Bangsa, Perlindungan dan Pemberdayaan Masyarakat (BKPPM) Kota Bandung, untuk sekitaran kota saja ada sekitar 6.000 warga yang menjadi bagian komunitas LGBT. Hal ini tentu menjadi suatu kekhawatiran di kalangan masyarakat Bandung yang notabene masih dikenal religius.

Gerakan LGBT berkembang cepat melalui media sosial, dengannya ide-ide LGBT dapat dengan leluasa dikemukakan. Esensi pesan berkenaan dengan pilihan hidup LGBT dapat tersampaikan, namun tanpa melibatkan eksistensi aslinya. Tweeter merupakan salah satu media yang seringkali digunakan untuk tujuan tersebut. Komentar-komentar atau "cuitan" tentang LGBT di twitter tentu banyak jumlahnya. Hartanto [3] pernah meneliti pengelompokkan komentar-komentar tentang LGBT di twitter, hasilnya terdapat 7 kelompok besar komentar. Adapun tujuan dari penelitian ini adalah mengelompokkan komentar-komentar di twitter tentang LGBT di Kota Bandung ke dalam *cluster-cluster* yang dapat dibedakan. Pengelompokan yang dihasilkan dapat menjadi petunjuk dasar terkait stigma masyarakat Bandung terkait berkembangnya LGBT. *Clustering* yang digunakan pada penelitian ini adalah K-*Means clustering*.

2. Tinjauan Pustaka

2.1 *Twitter*

Twitter didirikan oleh Jack Dorsey pada Maret 2006. Pada *platform* media sosial ini, pengguna tak terdaftar hanya bisa membaca kicauan sedangkan pengguna terdaftar bisa menulis kicauan melalui *Graphical User Interface* (GUI) situs, Pesan singkat (SMS), atau melalui berbagai aplikasi dari perangkat seluler.

Perkembangan jumlah pengguna Twitter meningkat dengan sangat cepat hingga dapat meraih popularitas di seluruh dunia. Hingga Januari 2013, tercatat sudah ada lebih dari 500 juta pengguna terdaftar. Popularitas Twitter yang makin meningkat menyebabkan layanan ini dimanfaatkan untuk berbagai keperluan diantaranya kampanye politik, pembentukan opini, sarana belajar, dan sebagai media komunikasi darurat. Twitter juga dihadapkan pada berbagai masalah dan kontroversi seperti masalah keamanan, pendidikan (bullying), dan privasi pengguna [3].

ISSN: 2527-3159 E-ISSN: 2527-3167

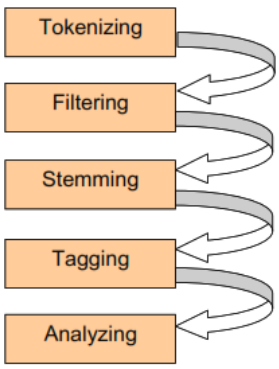
2.2 *Text Mining*

Banyaknya informasi yang ada di dunia maya membuat upaya-upaya pengembangan terhadap penggalian informasi dari basis data daring semakin pesat, salah satunya *text mining*. Text mining, yang juga disebut sebagai *Teks Data Mining* (TDM) atau *Knowledge Discovery in Text* (KDT), secara khusus dikembangkan untuk proses ekstraksi informasi dari dokumen-dokumen teks tak terstruktur *(unstructured)*. *Text mining* memiliki definisi menambang data berupa teks di mana sumber data biasanya didapatkan dari dokumen dan tujuanya adalah untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analis is keterhubungan antar dokumen [4].

Text mining mencoba memecahkan masalah kelebihan informasi *(information overload)* dengan menggunakan teknik-teknik dari bidang ilmu yang terkait. *Text mining* dapat dipandang sebagai perluasan dari *data mining* atau *Knowledge Discovery in Database* (KDD), yang bertujuan untuk menemukan pola- pola menarik dari basis data berskala besar.

Tahapan *Text Mining*

Tahapan *text mining* yang paling umum dilakukan adalah sebagai berikut



Gambar 1. Tahapan *Text Mining*

2.3 *Term Document Matrix (TDM)*

Term Document Matrix (tdm) adalah suatu matriks yang menggambarkan frekuensi kata yang terjadi dalam kumpulan dokumen. Dalam matriks tdm, banyaknya baris menggambarkan banyaknya dokumen sedangkan banyaknya kolom menggambarkan banyaknya kata [5].

2.4 *Term Frequency (TF)*

TF (*Term Frequency*) adalah frekuensi dari kemunculan sebuah istilah dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu term (TF tinggi) dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar [5].

2.5 IDF (*Inverse Document Frequency*)

IDF merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. IDF menunjukkan hubungan ketersediaan sebuah istilah dalam seluruh dokumen. Semakin sedikit jumlah dokumen yang mengandung term yang dimaksud, maka nilai IDF semakin besar**.** IDF dihitung dengan [5]:

$$IDF_j = \log (D / df_j) \quad (1)$$

di mana:
D : jumlah dokumen
df : jumlah dokumen yang mengandung term (tj).

selanjutnya untuk menghitung bobot (w) digunakan formula sebagai berikut:

$$W_{ij} = TF_{ij} \times IDF_j \quad (2)$$

di mana :
W_{ij} : adalah bobot term (t_j) terhadap dokumen (d_i)
TF_{ij} : jumlah kemunculan term (t_j) dalam dokumen (d_i)

2.6 *K-Means Cluster*

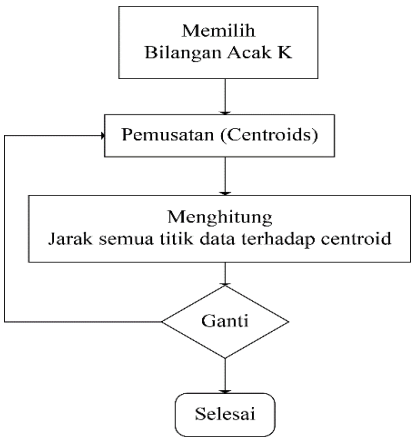
*Cluster*ing adalah salah satu metode yang dapat digunakan untuk mengeksplorasi distribusi dan pola data.

ISSN: 2527-3159 **E-ISSN: 2527-3167**

Pola-pola dalam suatu *cluster* akan memiliki kesamaan ciri/sifat dibandingkan pola-pola dalam *cluster* yang lainnya. *Cluster*ing bermanfaat untuk melakukan analisis pola-pola data, mengelompokkan, dan membuat keputusan . Ada beberapa algoritma dalam *cluster*ing, salah satu diantaranya adalah algoritma K-Means.

K-means ditemukan oleh beberapa orang yaitu Lloyd (1957, 1982), Forgey (1965), Friedman dan Rubin (1967), McQueen (1967) . Ide dari *clustering* pertama kali ditemukan oleh Lloyd pada tahun 1957, namun hal tersebut baru dipublikasi pada tahun 1982. Pada tahun 1965, Forgey juga mempublikasi teknik yang sama sehingga terkadang dikenal sebagai Lloyd-Forgy pada beberapa sumber.

Secara umum langkah-langkah dalam algoritma K-means *cluster*ing dapat diilustrasikan sebagai berikut [6]:



Gambar 2. algoritma K-means *cluster*ing

Berikut adalah algoritma dari metode k-means [8]:

a) Masukkan data yang akan diklaster.
b) Tentukan jumlah klaster.
c) Ambil sebarang data sebanyak jumlah klaster secara acak sebagai pusat klaster (sentroid).
d) Hitung jarak antara data dengan pusat klaster, dengan menggunakan persamaan :

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (3)$$

Dimana :
$D(i,j)$ = jarak data ke i ke pusat klaster j
X_{ki} = data ke i pada atribut ke k
X_{kj} = titik pusat ke j pada atribut ke k

e) Hitung kembali pusat klaster dengan keanggotaan klaster yang baru

Jika pusat klaster tidak berubah maka proses klaster telah selesai, jika belum maka ulangi langkah ke (d) sampai pusat klaster tidak berubah lagi.

3. Sumber Data

Data yang digunakan pada penelitian ini adalah data-data pada komentar-komentar yang ada di twitter yang mengandung kata LGBT di kota Bandung selama 10 hari yaitu pada tanggal 17 – 26 Desember 2017.

4. Analisis dan Pembahasan

4.1 Persiapan Data

Penelitian ini menggunakan program R. Jumlah dokumen/komentar yang diperoleh yaitu sebanyak 701 dokumen/komentar. Tentu saja data yang diperoleh tidak bisa langsung digunakan, akan tetapi harus melalui tahapan cleaning data dengan menggunakan tahapan-tahapan text mining yang telah disebutkan pada pembahasan sebelumnya yaitu tokenizing, filtering dan stemming. Tahapan tagging tidak dilakukan karena tahapan ini hanya bisa dilakukan pada dokumen yang berbahasa inggris. Setelah melalui tahapan cleaning data diperoleh sebanyak 691 dokumen yang siap untuk dianalisis menggunakan K-means *cluster*ing.

4.2 *Term Document Matrix* TDM dan *Term Frequency-Invers Document Frequency*

Dari proses yang dilakukan diperoleh matriks tdm berikut:

ISSN: 2527-3159 E-ISSN: 2527-3167

	agama	agamanya	aja	apapun	kecuali	kitab	melarang	...
1	1	1	1	1	1	1	1	...
2	0	0	0	0	0	0	0	...
3	1	0	0	0	0	0	0	...
4	0	0	0	0	0	0	0	...
5	0	0	0	0	0	0	0	...
6	0	0	0	0	0	0	0	...
7	0	0	0	0	0	0	0	...
8	0	0	0	0	0	0	0	...
9	0	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...

Gambar 3. Matriks TDM

Dan matriks pembobotan **TF-IDF**

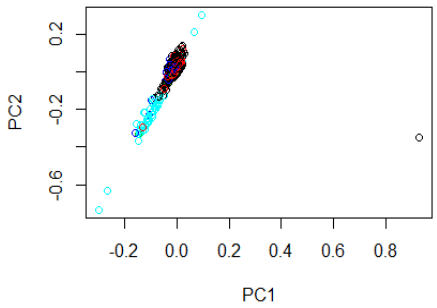
	agama	agamanya	aja	apapun	kecuali	kitab	melarang	...
1	0.4736	1.0540	0.5050	0.8888	0.9290	1.0541	0.7638	...
2	0	0	0	0	0	0	0	...
3	0.9472	0	0	0	0	0	0	...
4	0	0	0	0	0	0	0	...
5	0	0	0	0	0	0	0	...
6	0	0	0	0	0	0	0	...
7	0	0	0	0	0	0	0	...
8	0	0	0	0	0	0	0	...
9	0	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...

Gambar 4. Matriks bobot TF-IDF

4.3 K-Means *Cluster*

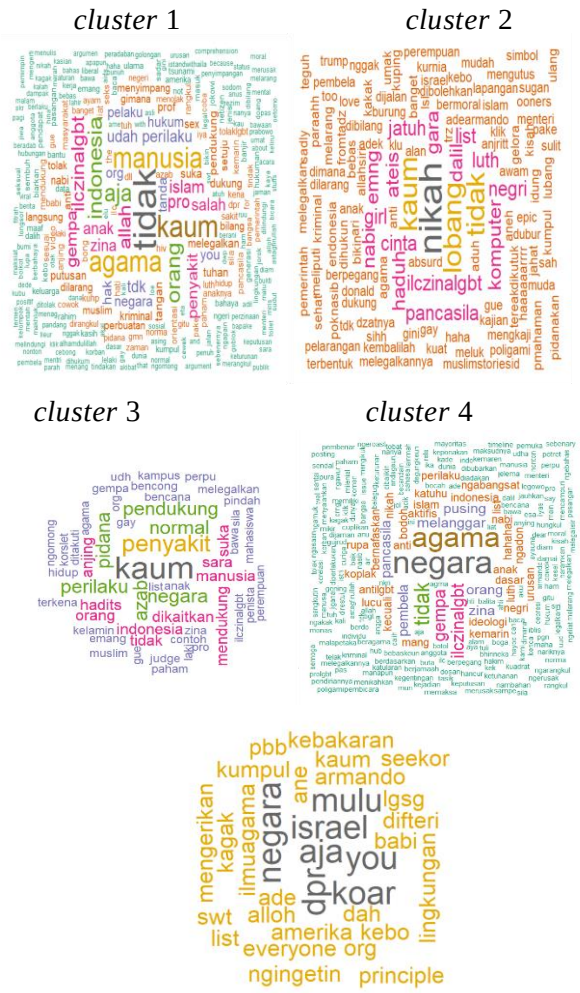
Penelitian ini menggunakan nilai K=5. Dari lima *cluster* yang ditentukan tersebut setelah dilakukan pemrosesan dengan paket program R diperoleh informasi bahwa *cluster* satu yang terbentuk beranggotakan 51 sampel dokumen. *Cluster* dua diperoleh 89 sampel, *cluster* tiga terdapat 80 sampel, *cluster* empat 419 sampel, dan *cluster* lima 34 sampel. Informasi lain yang diperoleh adalah jumlah kuadrat (sum square) pada masing-masing *cluster*. Jumlah kuadrat dalam *cluster* di *cluster* satu sebesar 45,069; *cluster* dua sebesar 84,64; *cluster* tiga 75,781; *cluster* empat 414,736; dan *cluster* lima 30,175.

Ada kesulitan yang dihadapi dalam menganalisis text berbahasa Indonesia. Corpus bahasa Indonesia yang lengkap masih sulit diperoleh selain itu seringkali komentar-komentar atau cutan-cuitan dalam Twitter tidak menggunakan bahasa Indonesia baku. Kendala ini menyebabkan hasil yang diperoleh dari data mining yang dilakukan mesin masih belum bisa seakurat yang diinginkan. Jika dibuat plot, pengelompokan yang terbentuk sebagai berikut:



Gambar 5. Plot *Cluster* K-Means

Berikutnya dapat dilihat pola sebaran kata dalam setiap *cluster* dengan menggunakan wordcloud di bawah ini:



Gambar 6. Wordcloud pada *cluster*

ISSN: 2527-3159 **E-ISSN: 2527-3167**

Dari plot-plot tersebut dapat dilihat bahwa *centroid* pada masing-masing *cluster* berbeda-beda. Pusat perhatian *cluster* satu pada kata manusia, agama, dan kaum; *cluster* kedua lebih pada kata nikah, dan kata lobang; *cluster* tiga kata kaum, adzab dan penyakit; *cluster* empat pada kata agama dan negara; dan yang terakhir dihubungkan dengan kata israel dan DPR. Untuk mendapatkan gambaran yang lebih jelas tentang isi dari masing- masing dapat pula dikaji beberapa dokumen yang masuk dalam suatu *cluster*. Misalnya, untuk *cluster* satu jika dikaji lebih dalam didapati bahwa artikel-artikel di dalamnya lebih banyak berbicara tentang hukum perilaku LGBT menurut sudut pandang agama Islam. *Cluster* ke-dua lebih kepada masalah pembahasan moral dan *cluster* 3 pada opini tentang dampak perilaku LGBT.

5. Kesimpulan

Dari tahapan-tahapan yang telah dilewati diperoleh gambaran bahwa penggunaan K-mean *clustering* dapat digunakan untuk pembentukan *cluster* kata pada arsip-arsip dokumen yang digunakan. Sayangnya, kendala yang dihadapi pada saat proses text mining yang tidak sempurna memfilter kata-kata dalam bahasa Indonesia yang digunakan pengguna *Tweeter* dalam arsip komentar menyebabkan *cluster* yang terbentuk masing mengandung kata-kata yang tidak begitu penting.

Dari lima *cluster* yang dibentuk pada proses K-means diperoleh bahwa kecenderungan cuitan pengguna *Tweeter* kota bandung terkait LGBT secara umum masih berhubungan dangan perspektif religi. Kemunculan kata agama yang sangat sering menyebabkan asosiasi terhadap kata tersebut cukup besar.

Referensi

[1] Sinyo, Anakku Bertanya Tentang LGBT, PT Elex Media Komputindo, (2014).
[2] Gatra, Melawan Aksi LGBT di Kampus, (2016).
[3] Alim, S, Analysis of Tweets Related to Cyberbullying: Exploring Information Diffusion and Advice Available for Cyberbullying Victims. International
[4] Journal of Cyber Behavior, Psychology and Learning, (2015).
[5] Prasetyo, Eko, Data Mining Konsep dan Aplikasi menggunakan Matlab, Penerbit Andi Yogyakarta (2012).
[6] Srihari, Retrieval by Content, diambil darihttp://www.cedar.buffalo.edu/~srihari/ CSE626/Lecture-Slides, pada tanggal 6 Februari 2018
[7] Hastuti, N. F., Saptono, R., &Suryani, E., Pemanfaatan Metode K-Means Clustering Dalam Penentuan Penerima Beasiswa,Jurnal Informatika, (2012).
[8] Febrianti, F., Hafiyusholeh, M., & Asyhar, A.H., Perbandingan Pengklusteran Data Iris Menggunakan Metode K-Means dan Fuzzy C-Means, Jurnal Matematika MANTIK, 2 (1), pp. 7-13 (2016)