PENGGEROMBOLAN TWEET BADAN NASIONAL PENANGGULANGAN BENCANA INDONESIA PERIODE AGUSTUS 2018–FEBRUARI 2019 MENGGUNAKAN *TEXT MINING*[*]

Windyana Pusparani[1], Agus M Soleh[2‡], and Akbar Rizki[3]

[1]Department of Statistics, IPB University, Indonesia, windyana_puspa@apps.ipb.ac.id
[2]Department of Statistics, IPB University, Indonesia, agusms@apps.ipb.ac.id
[3]Department of Statistics, IPB University, Indonesia, akbar.ritzki@gmail.com
[‡]corresponding author

Indonesian Journal of Statistics and Its Applications (eISSN:2599-0802)
Vol 4 No 4 (2020), 590 - 603

Abstract

Twitter is a popular social media platform for communicating between its users by writing short messages in limited characters, called tweets. Extracting data information that has non-structured form and huge-sized, usually known as text mining. Badan Nasional Penanggulangan Bencana Indonesia (@BNPB_Indonesia) is the official twitter account of the government agency in the field of disaster management that uses twitter to share much information about disasters that have occurred in Indonesia. This study aims to determine the characteristics of all tweets and to group the types of tweets that they shared based on the similarity of its content. The data used in the study came from BNPB Indonesia's tweets with the period of taking tweets 6^{th} of August 2018 to 16^{th} of February 2019. The cluster result obtained by the k-Means method was 4 groups. The characteristics of the first cluster contained information about the weather conditions in Yogyakarta, the second cluster was about the source and magnitude of an earthquake, and the third group was about the occurrence of earthquakes in Lombok. However, the fourth group characteristic couldn't be specifically identified because there was no clear distinction between other tweets in its members.

Keywords: clustering analysis, disaster, k-Means, text mining.

[*]Received Aug 2019; Accepted Des 2020; Published online on Des 2020

1 Pendahuluan

Hasil survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) tahun 2017 menyatakan bahwa, pengguna internet di Indonesia diduga mencapai 143.26 juta jiwa atau 54.68% dari total penduduk Indonesia. Sebanyak 87.13% dari 54.68% ini menggunakan internet untuk mengakses media sosial. *Twitter* menjadi salah satu *platform* media sosial yang populer dalam kegiatan *microblogging*, yang memungkinkan antar pengguna untuk saling menuliskan *tweet*.

Bencana alam yang terjadi secara beruntun di Indonesia pada akhir pertengahan tahun 2018 hingga awal tahun 2019 menjadi salah satu topik perbincangan yang ramai dibicarakan oleh para pengguna *twitter*. Akun *twitter* BNPB Indonesia (@BNPB_Indonesia) merupakan akun resmi lembaga pemerintahan yang secara aktif menuliskan *tweet* untuk menyebarkan informasi mengenai penanganan dampak bencana yang terjadi di Indonesia. Pentingnya peran akun *twitter* BNPB sebagai penyampai informasi bencana membuat perlunya dilakukan penelitian untuk mengelompokan isi *tweet* dari akun ini. Hal ini dapat digunakan sebagai kontrol apakah akun *twitter* BNPB telah menjalankan perannya dengan baik.

Setiap satu *tweet* berbentuk teks yang dibagikan oleh sebuah akun *twitter* akan menghasilkan satu dokumen, yang jika dikumpulkan dan diolah dengan metode tertentu dapat menghasilkan informasi yang berguna. *Text mining* diartikan sebagai proses untuk mendapatkan informasi dari berbagai dokumen dalam bentuk teks yang dapat bersumber dari sebuah *e-mail*, makalah penelitian, hingga kabar berita (Feldman *et al.*, 2007). (Chen *et al.*, 2014) menerapkan text mining dengan menggunakan percakapan dalam media sosial untuk memahami masalah mahasiswa jurusan teknik dalam kehidupan perkuliahan seperti beban studi yang berat, kekurangan keterlibatan sosial dan waktu untuk beristirahat. (Slamet *et al.*, 2016) melakukan penggerombolan data teks berupa penggerombolan ayat-ayat suci Al-Qur'an untuk mengetahui kelompok konten dan eksplorasi ilmu pengetahuan yang terdapat dalam Al-Qur'an menggunakan metode k-*Means*.

Penelitian ini berupaya untuk menggali informasi dari data tekstual *tweet* akun *twitter* BNPB untuk mengetahui kelompok-kelompok topik pembicaraan yang disampaikan akun tersebut menggunakan metode k-*Means*. Keuntungan dari penggunaan metode k-*Means* untuk penggerombolan teks adalah tidak memerlukan iterasi yang banyak untuk mendapatkan hasil yang baik, sehingga tepat untuk diterapkan pada data teks yang berukuran besar (Mattjik & Sumertajaya, 2011). Karakteristik setiap kelompok topik kemudian akan diidentifikasi berdasarkan hasil visualisasi teksnya.

2 Metodologi

2.1 Data

Data penelitian yang digunakan adalah data *tweet* pada akun media sosial twitter milik Badan Nasional Penanggulangan Bencana (BNPB) Indonesia(@BNPB_Indonesia), baik *tweet* yang dibuat oleh BNPB sendiri maupun yang berasal dari me*retweet tweet* dari akun *twitter* pengguna/lembaga lain. *Tweet* yang terambil adalah 3000 *tweet* teratas sejak tanggal 6 Agustus 2018 hingga 16 Februari 2019. Data *tweet* didapatkan setelah mendaftarkan diri melalui *twitter* REST API yang merupakan sistem antarmuka dalam media sosial untuk mendapatkan data dengan format tertentu.

2.2 Prosedur Analisis Data

1. Eksplorasi Data

   (a) Membuat hasil eksplorasi data berupa diagram tipe *tweet*, kategori *tweet*, dan kategori bencana.

   (b) Menyajikan 10 kata yang paling sering muncul dengan frekuensi kemunculannya dan hasil visualisasi data tekstual berupa *word cloud*.

2. Pra proses data

   (a) Pra proses tahap 1: Mengubah huruf kapital menjadi nonkapital (*lowercase*), memisahkan kalimat teks ke dalam individu kata dan frasa, menghapus tanda baca, URL, *mentions* ("RT"), *username*, angka, dan *hashtag*.

   (b) Pra proses tahap 2: Membuat daftar *stopwords* berdasarkan pengamatan pada seluruh teks yang dimiliki dan menggabungkannya dengan *stopwords* bahasa Indonesia dari penelitian Tala (2003), dan *stopwords* bahasa Inggris dari *Natural Language Toolkit*. Bila *stopwords* ditemukan dalam dokumen, maka akan dihilangkan.

   (c) Pra proses tahap 3: Mengubah teks ke bentuk dasarnya (*stemming*) berdasarkan algoritma *confix-stripping* Adriani *et al.* (2007) dengan menghapus awalan, imbuhan, dan akhiran dari setiap kata yang ditemukan.

3. Transformasi Data

   (a) Membuat suatu *corpus* dokumen dari data *tweet* yang telah melalui seluruh tahap pra proses.

   (b) Membentuk *Document Term Matrix* yaitu matriks berukuran n *x* p dengan n adalah banyaknya dokumen sebanyak 3000, p adalah banyaknya kata sebanyak 1402, yang berisi pembobotan TF-IDF terhadap seluruh kata di seluruh *tweet*. Pembobotan kata TF-IDF untuk kata ke$-t$ dalam dokumen ke$-d$ adalah sebagai berikut:

   $$Tfidf_{td} = Tf_{td} \times \log \frac{D}{df_d}$$

   dalam hal ini $Tfidf_{td}$ adalah bobot kata ke$-t$ dalam dokumen ke$-d$, Tf_{td} adalah frekuensi kemunculan kata ke$-t$ pada dokumen ke$-d$ dibagi banyaknya kata dalam dokumen ke$-d$, D adalah banyaknya dokumen, dan df_d adalah banyaknya dokumen yang mengandung kata ke$-t$ (Salton & Buckley, 1988).

   (c) Menghitung nilai *cosine similarity* dari *Document Term Matrix* yaitu matriks berukuran n *x* n atau 3000 *x* 3000. *Cosine similarity* adalah sebuah ukuran kemiripan dokumen dengan menghitung nilai *cosine* dari sudut yang dibentuk oleh dua buah vektor. Misalkan $\mathbf{x}$ dan $\mathbf{y}$ adalah dua vektor pembobotan kata dengan TF-IDF, maka ukuran kemiripan ini dihitung dengan rumus,

   $$sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^t\mathbf{y}}{||\mathbf{x}||||\mathbf{y}||}$$

   bila nilai *cosine similarity* bernilai 0, berarti sudut antara kedua vektor 90^o , $\mathbf{x}$ dan $\mathbf{y}$ tidak memiliki kemiripan. Semakin dekat nilai *cosine* dengan 1, maka semakin kecil sudut yang dibentuk vektor $\mathbf{x}$ dan $\mathbf{y}$ dan semakin mirip kedua vektor atau dokumen tersebut.

4. Penggerombolan Data dengan k-*Means*

   (a) Menggunakan matriks *cosine similarity* sebagai data yang akan digerombolkan dengan k-*Means*.

   (b) Menggerombolkan data dengan banyaknya gerombol k-*Means* yang dicobakan yaitu k=2 hingga k=7.

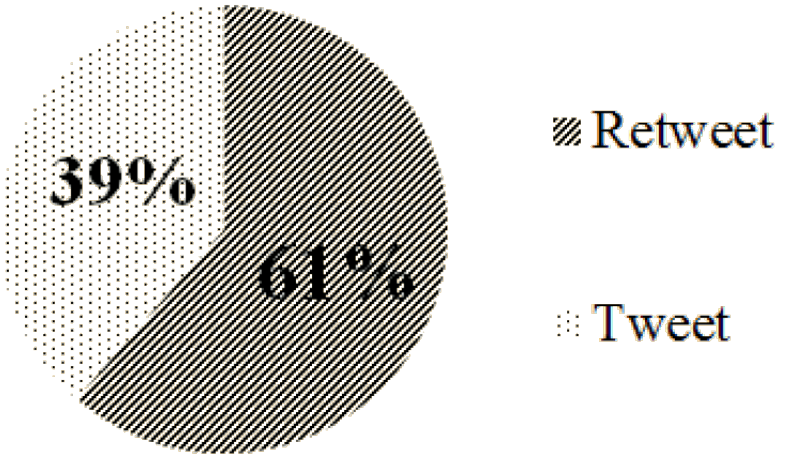
   (c) Menentukan (k) terpilih sebagai panduan penggerombolan berdasarkan hasil perhitungan nilai koefisien *silhouette* tertinggi.

   (d) Mengidentifikasi karakteristik hasil gerombol awal dengan melihat visualisasi *word cloud* anggota masing-masing gerombol dan melakukan *overfitting* bila karakteristik gerombol yang dihasilkan terlihat masih dapat dipisahkan menjadi gerombol lainnya.

3 Hasil dan Pembahasan

3.1 Eksplorasi Data

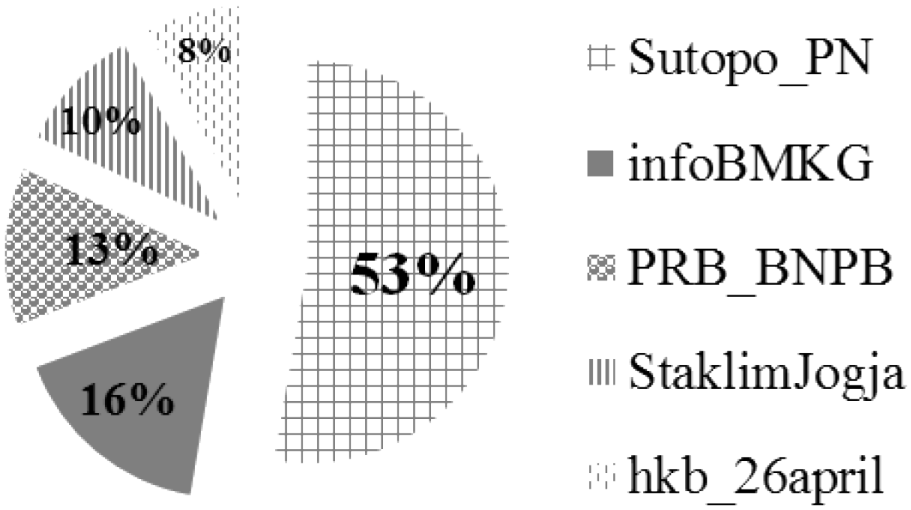
Gambar 1 menyatakan proporsi antara banyaknya *tweet* yang di*tweet* oleh akun @BNPB_Indonesia dan yang di*retweet* dari pengguna *twitter* lain. Besar proporsi ini menunjukkan bahwa *tweet* hasil *retweet* lebih banyak dibandingkan *tweet* yang ditulis sendiri. Sebanyak 61% atau 1842 dari 3000 *tweet* yang didapatkan berasal dari *tweet* hasil *retweet* dari akun *twitter* lain, sedangkan 39% atau 1158 data adalah *tweet* yang dituliskan sendiri.



Gambar 1: Proporsi tipe *tweet*

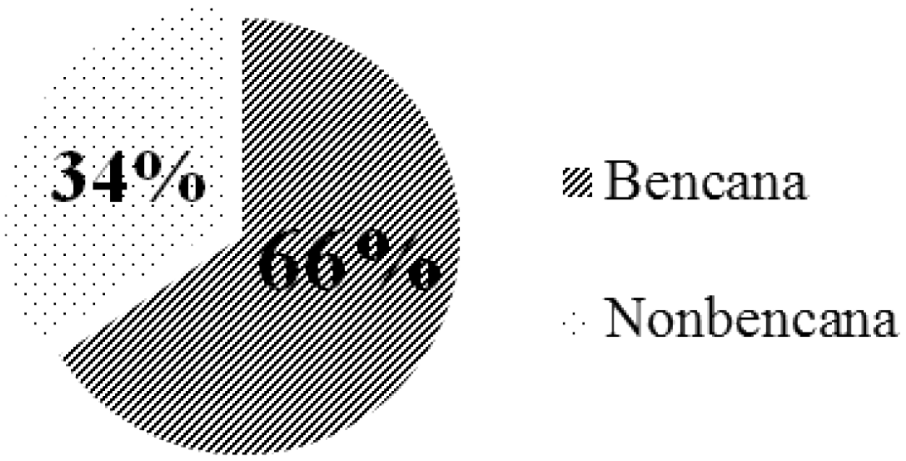
Tweet hasil *retweet* ini didapatkan dari 246 akun berbeda, dengan lima akun paling sering di*retweet* adalah milik @Sutopo_PN, @infoBMKG, @PRB_BNPB, @StaklimJogja, dan @hkb_26april. Presentasi frekuensi *retweet* total dari lima akun tersebut sebesar 39% dan proporsi masing-masing disediakan dalam Gambar 2. Akun yang *tweet*nya paling sering di*retweet* adalah akun @Sutopo_PN, milik Sutopo Purwo Nugroho yang menjabat Kepala Pusat Data Informasi dan Humas BNPB Indonesia. Akun urutan

kedua yaitu akun @infoBMKG, milik Badan Meteorologi, Klimatologi, dan Geofisika yang menuliskan informasi mengenai cuaca, iklim, maupun peringatan dini bencana. Akun ketiga adalah akun @PRB_BNPB, milik Direktorat Pengurangan Risiko Bencana BNPB. Akun keempat yaitu @StaklimJogja, milik Stasiun Klimatologi Mlati Yogyakarta yang menyebarkan informasi mengenai prakiraan cuaca di Yogyakarta. Akun kelima adalah @hkb_26april, akun milik panitia peringatan Hari Kesiapsiagaan Bencana yang dibentuk oleh Direktorat Kesiapsiagaan BNPB Indonesia.



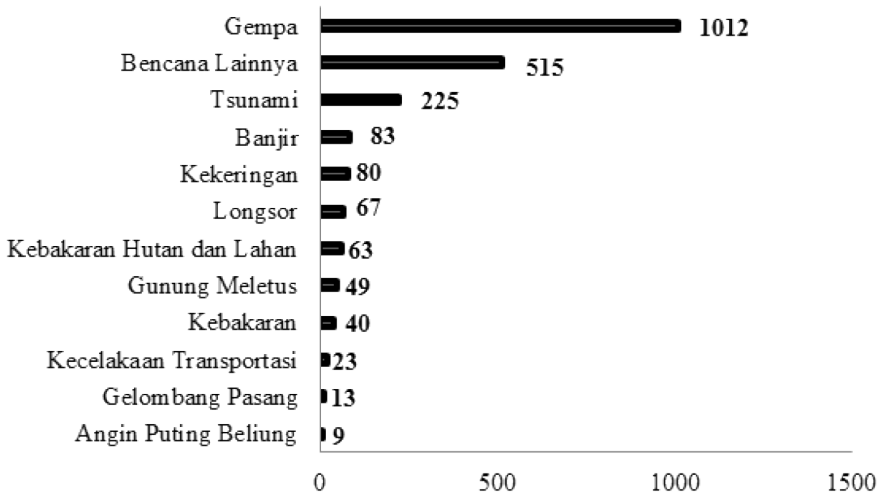
Gambar 2: Proporsi akun yang di*retweet*

Gambar 3 menyatakan proporsi kategori *tweet* yaitu bencana dan nonbencana dengan *tweet* bencana lebih banyak terdapat dalam data dibandingkan *tweet* mengenai nonbencana. Sebanyak 66% atau 1975 *tweet* membicarakan bencana, sedangkan sisanya sebesar 34% atau 1025 *tweet* mengenai nonbencana. Hal ini memperlihatkan bahwa akun @BNPB_Indonesia tidak hanya menuliskan *tweet* atau me*retweet* hal yang berkaitan dengan bencana saja, namun juga hal selain bencana.



Gambar 3: Proporsi kategori *tweet*

Terdapat 18 kategori bencana menurut BNPB Indonesia yang kemudian disesuaikan dengan seluruh *tweet* yang dimiliki sehingga menghasilkan 13 kategori. Gambar 4 menyajikan frekuensi masing-masing kategori bencana, terlihat bahwa bencana yang paling sering disebut adalah gempa bumi sebanyak 1012 kali. Menurut publikasi infografis kejadian bencana yang diterbitkan di laman *website* BNPB, sepanjang Agustus 2018 hingga Februari 2019 gempa bumi telah terjadi sebanyak 20 kali kejadian yang tercatat pertama kali terjadi pada 5 Agustus 2018 berkekuatan 6.4 SR di Lombok. Gempa bumi menjadi hal paling sering dibahas karena bencana yang menelan ribuan korban jiwa ini berturut-turut mengguncang pulau-pulau lainnya di Indonesia seperti Palu dan Donggala, Bali, dan Nusa Tenggara Timur.

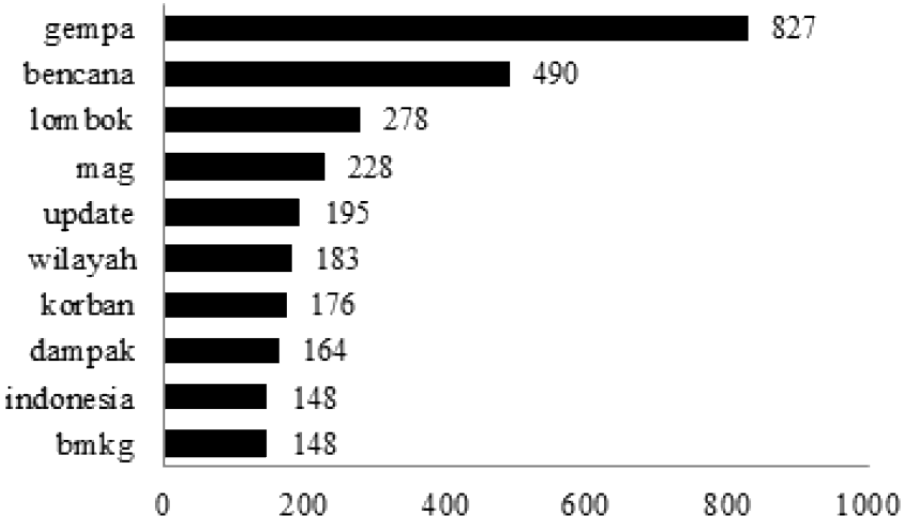


Gambar 4: Grafik kategori bencana

Terdapat sepuluh kata dengan frekuensi kemunculan terbanyak dalam seluruh *tweet* yang berhubungan dengan bencana dengan rincian seperti Gambar 5. Hal ini sudah sesuai dengan peran akun *twitter* @BNPB_Indonesia yang bertugas menyebarkan informasi mengenai bencana yang terjadi di Indonesia. Kata "gempa" sebagai kata terbanyak pertama yang ditemukan menandai banyaknya peristiwa bencana gempa bumi yang terjadi selama periode pengambilan *tweet* dalam penelitian ini.

Frekuensi kemunculan seluruh kata dalam keseluruhan *tweet* kemudian disajikan dalam *word cloud* seperti disajikan pada Gambar 6. Kata dengan ukuran terbesar dapat dikatakan adalah kata yang menjadi topik pembicaraan dalam seluruh *tweet*. Hasil visualisasi pada Gambar 6 telah sejalan dengan diagram batang pada Gambar 5, bahwa 10 kata terbanyak yang muncul terlihat jelas pada Gambar 6.

Tabel 1 menyajikan ilustrasi perubahan *tweet* awal menjadi *tweet* yang telah melalui seluruh tahap praproses. Hasil pra proses tahapan pertama adalah seluruh *tweet* telah menggunakan huruf nonkapital dan seluruh tanda baca telah dihilangkan. Penghapusan link Uniform Resource Locator (URL), *mention* ("RT") , *username* akun, angka, dan *hashtag*(#) yang tidak diikutsertakan dalam dianalisis. Hasil pra proses tahapan kedua adalah seluruh dokumen yang stopwordsnya telah dihilangkan berdasarkan daftar stopwords. Contoh penghapusan stopwords pada ilustrasi Tabel 1 adalah penghilangan kata "dini", "di", "tanggal", "februari", "pukul", dan "wib" dalam *tweet* pertama. Penghapusan kata "pgr", "info", "sr", "lok", "ls", "bt", "kedlmn", dan "km" pada *tweet* kedua. Penghilangan kata "dalam" dan "letnan" dari *tweet* ketiga. Hasil pra proses



Gambar 5: Kata dengan frekuensi kemunculan terbanyak

Gambar 6: *Word cloud* seluruh *tweet*

tahapan terakhir adalah seluruh dokumen yang hanya memiliki kata dasar saja. Misalkan terdapat kata "peringatan", "mengingatkan", dan "diperingati" diubah ke kata dasarnya yaitu "ingat", sehingga kata-kata tersebut menjadi satu makna dan selanjutnya dapat memudahkan dalam analisis. Tabel 1 memperlihatkan pengubahan kata "peringatan" menjadi "ingat" pada *tweet* pertama dan kata "pendukung" menjadi "dukung" pada *tweet* ketiga.

Tabel 1: Ilustrasi perubahan *tweet* awal menjadi *tweet* yang telah melalui tahap praproses

Tweet ke-	*Tweet* Awal	Setelah Pra Proses
1	RT @StaklimJogja: Update Peringatan Dini Cuaca Wilayah DI. Yogyakarta Tanggal 16 Februari 2019 pukul 15.10 WIB #InfoCuacaJogja #BMKGDIY	update ingat cuaca wilayah yogyakarta infocuacajogja bmkgdiy
2	PGR 3 Bali: Info Gempa Mag:3.2 SR, 16-Feb-19 15:06:13 WIB, Lok:9.65 LS,113.97 BT (160 km BaratDaya JEMBRANA-BALI), Kedlmn:19 Km ::BMKG	bali gempa mag baratdaya jembrana bali bmkg
3	TNI Komponen Pendukung dalam Penanggulangan Bencana. JAKARTA - Kepala Badan Nasional Penanggulangan Bencana Letnan ... https://t.co/iWT1FHqnsC	tni komponen dukung penanggulangan bencana jakarta badan nasional penanggulangan bencana

3.2 **Transformasi Data**

Ilustrasi pembuatan *Document Term Matrix* dapat dilihat pada Tabel 2 yaitu matriks berukuran 3000 x 1402. Setiap kata yang terdapat pada *tweet* akan menjadi suatu peubah atau kolom dari matriks, yang berisi nilai seluruh kata di suatu *tweet* tertentu yang dihitung berdasarkan pembobotan kata TF-IDF.

Tabel 2: Ilustrasi pembuatan *document term matrix* berdasarkan TF-IDF

Dokumen ke-	giat	identifikasi	sosial	...	runtuh
	1	2	3	...	1402
1	0.49	0.49	0.41	...	0
2	0	0	0	...	0
3	0	0	0	...	0
...	...	...	...	...	...
2999	0	0	0	...	0
3000	0	0	0	...	0

Tabel 3 merupakan contoh tampilan *cosine similarity matrix* yaitu matriks yang berukuran n x n atau 3000 x 3000 dari *Document Term Matrix* yang berasal dari jarak kemiripan *cosine* antar satu dokumen dengan dokumen-dokumen lainnya. Nilai-nilai yang ada dalam matriks *cosine similarity* akan digunakan sebagai dasar pengelompokkan data dengan metode k-*Means*.

Tabel 3: *Cosine Similarity Matrix*

	Dokumen1	Dokumen2	Dokumen3	Dokumen4	Dokumen5
Dokumen1	1	0	0	0.245	0
Dokumen2	0	1	0	0	0
Dokumen3	0	0	1	0	0
Dokumen4	0.245	0	0	1	0
Dokumen5	0	0	0	0	1

3.3 Hasil Penggerombolan

Tabel 4 menyajikan daftar nilai koefisien *silhouette* untuk banyaknya gerombol sebanyak 2 hingga 7 gerombol. Terlihat bahwa nilai koefisien silhouette tertinggi yaitu sebesar 0.338 dengan banyaknya gerombol sebanyak 3 gerombol. Banyaknya gerombol ini akan dijadikan panduan banyaknya gerombol awal.

Tabel 4: Nilai koefisien *silhouette* pada k=2 hingga k=7

Banyaknya Gerombol (k)	Koefisien *Silhouette*
2	0.313
3	**0.338**
4	0.164
5	0.174
6	0.162
7	0.107

3.3.1 Karakteristik Hasil 3 Gerombol

Banyaknya anggota masing-masing gerombol pada penggerombolan *tweet* ke dalam 3 gerombol disajikan pada Tabel 5. Gerombol pertama banyaknya anggota paling sedikit yaitu 56 *tweet*. Gerombol 3 berisi anggota terbanyak yaitu 2710 *tweet* yang menunjukkan bahwa tidak ditemukannya kata yang dominan sebagai pembeda karakteristiknya dengan gerombol yang lain.

Tabel 5: Banyaknya anggota masing-masing gerombol

Gerombol ke-	Banyaknya Anggota
1	56
2	234
3	2710

Gambar 7 menyajikan visualisasi anggota Gerombol 1. Terlihat bahwa data yang termasuk dalam anggota gerombol pertama adalah *tweet* mengenai *update* keadaan cuaca di Yogyakarta berdasarkan pantauan citra radar BMKG Yogyakarta. Hal ini mengindikasikan *tweet* hasil *retweet* dari akun @StaklimJogja sebagai akun peringkat keempat paling sering di*retweet* akun BNPB Indonesia, yang menyebarkan *update*

informasi cuaca Yogyakarta kemudian membentuk sebuah gerombol yang berdiri sendiri. Contoh salah satu *tweet* anggota gerombol ini adalah "RT @StaklimJogja: Update Citra Radar Cuaca Wilayah DI. Yogyakarta Tanggal 17 Maret 2019 pukul 05.00 WIB #InfoCuacaJogja #BMKGDIY".

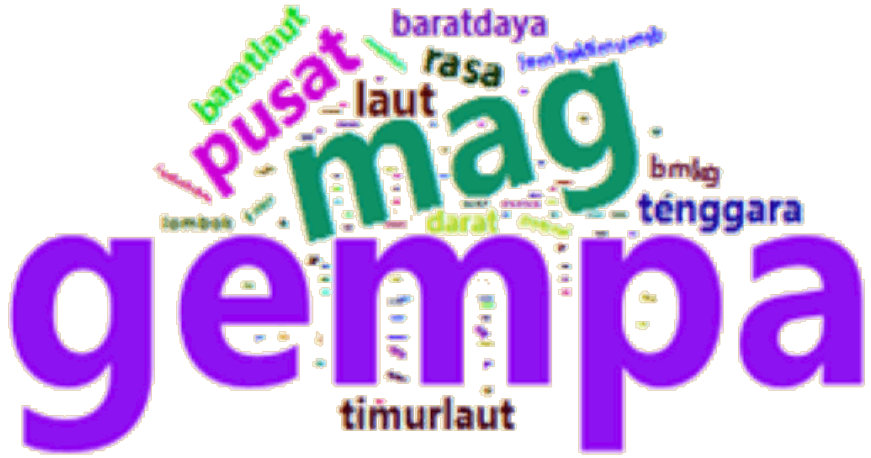
Gambar 7: *Word cloud* anggota Gerombol 1

Anggota gerombol kedua berdasarkan Gambar 8 adalah kumpulan *tweet* yang memberi informasi mengenai sumber pusat gempa bumi dan nilai besarannya. Sumber gempa bumi didefinisikan contohnya oleh kata "baratdaya", "tenggara" yang menunjukkan pusat gempa terjadi, sementara nilai besaran gempa bumi ditandai dengan kata "mag" yang berarti "magnitude". Hal ini mengindikasikan *tweet* hasil *retweet* dari akun @infoBMKG sebagai akun urutan kedua paling sering di*retweet* akun BNPB Indonesia yang membicarakan pusat dan besaran gempa membentuk suatu gerombol dengan salah satu *tweet*nya adalah "RT @infoBMKG: #Gempa Mag:5.6, 13-Okt-18 11:34:16 WIB, Lok:1.36 LU, 125.46 BT (Pusat gempa berada di Laut 38 km Tenggara Bitung), Kedlmn:97".

Gambar 9 menyajikan visualisasi anggota gerombol terakhir yaitu *tweet* yang membicarakan beragam hal dan tidak ada kecenderungannya terhadap topik tertentu. Hal ini menunjukkan bahwa penggerombolan pada Gerombol 3 belum terpisah dengan baik karena tidak ditemukan pembeda antar satu *tweet* dengan *tweet* yang ada dalam gerombolnya sehingga gerombol tidak membicarakan hal tertentu yang dapat diidentifikasi jenis kontennya. Karakteristik yang belum spesifik inilah yang menyebabkan dilakukannya *overfitting* menjadi 4 gerombol.

3.3.2 Karakteristik Hasil 4 Gerombol

Tabel 6 menyajikan banyaknya anggota hasil gerombol sebanyak 4. Terlihat bahwa banyaknya anggota gerombol pertama dan kedua tidak jauh berbeda dari penggerombolan



Gambar 8: *Word cloud* anggota Gerombol 2

Gambar 9: *Word cloud* anggota Gerombol 3

sebelumnya, namun Gerombol 3 terpisah menjadi gerombol ketiga dan keempat. Terlihat

Tabel 6: Banyaknya anggota masing-masing gerombol

Gerombol ke-	Banyaknya Anggota
1	56
2	222
3	496
4	2226

pada visualisasi anggota Gerombol 1 dan 2 pada Gambar 10, bahwa data yang termasuk dalam anggota gerombol pertama dan kedua sama dengan anggota gerombol pertama dan kedua pada banyaknya 3 gerombol. Hal ini menandakan kekonsistenan anggota gerombol pertama dan kedua, sehingga penggerombolan keduanya sudah baik.

Gambar 10: *Word cloud* anggota Gerombol 1 (a) dan (b) Gerombol 2

Gambar 11 menyajikan visualisasi anggota Gerombol 3. Anggota gerombol ketiga adalah *tweet* mengenai bencana gempa bumi di Lombok pada 5 Agustus 2018 yang diwakili kata "gempa" dan "lombok" sebagai kata yang mendominasi anggota Gerombol 3. Meskipun jenis konten yang disimpulkan adalah kejadian gempa bumi di Lombok, anggota Gerombol 3 juga bercampur dengan peristiwa gempa yang diikuti tsunami setelahnya di Palu dan Donggala, Sulawesi Tengah pada 28 September 2018 yang ditunjukkan oleh kata "tsunami" dan "sulteng" pada *word cloud*.

Terlihat pada Gambar 12 bahwa anggota gerombol keempat adalah *tweet* yang membicarakan beragam hal dan tidak spesifik kecenderungannya terhadap topik tertentu. Hal ini menunjukkan bahwa penggerombolan pada gerombol 4 tidak terpisah dengan baik karena tidak ditemukan pembeda antar satu *tweet* dengan *tweet* lainnya yang ada dalam gerombolnya sehingga tidak dapat diidentifikasi jenis kontennya.

Hasil penggerombolan k-*Means* yang sudah dilakukan melalui proses pemilihan banyaknya gerombol berdasarkan nilai *silhouette coefficient* kemudian dilakukannya proses *overfitting* menunjukkan bahwa gerombol terbaik adalah gerombol sebanyak 4. Hal ini diperkuat dengan hasil penentuan konten anggota berdasarkan karakteristiknya, terlihat bahwa gerombol pertama dan kedua pada penggerombolan ke dalam 3 maupun

Gambar 11: *Word cloud* anggota Gerombol 3

Gambar 12: *Word cloud* anggota Gerombol 4

4 gerombol konsisten membicarakan mengenai gempa bumi dan keadaan cuaca Yogyakarta. Perlakuan *overfitting* untuk memisahkan anggota gerombol ketiga menjadi gerombol ketiga dan keempat diketahui telah mempengaruhi hasil penentuan karakteristik sebelumnya karena pembeda dari karakteristik konten telah dapat didefinisikan dengan jelas.

4 Simpulan dan Saran

Penggerombolan *tweet* pada akun *twitter* BNPB Indonesia (@BNPB Indonesia) menggunakan metode *text mining* dan analisis gerombol tak berhierarki k-Means berdasarkan kemiripan karakteristiknya menghasilkan gerombol terbaik sebanyak 4. Hal ini didasarkan pada hasil *overfitting* dari penentuan banyaknya gerombol menurut koefisien *silhouette* tertinggi. Karakteristik anggota Gerombol 1 adalah *tweet* mengenai *update* cuaca Yogyakarta melalui citra radar, Gerombol 2 tentang sumber dan besaran gempa bumi, serta Gerombol 3 tentang kejadian gempa bumi di Lombok. Namun demikian, Gerombol 4 tidak dapat diidentifikasi karakteristiknya karena tidak ditemukan pembeda yang jelas antar *tweet* dalam gerombolnya.

Penelitian selanjutnya pada penggerombolan *tweet* dapat menggunakan metode *spherical* k-*means* yang bermanfaat untuk membandingkan dan melihat metode yang memiliki hasil penggerombolan lebih baik.

Daftar Pustaka

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M., & Williams, H. E. (2007). Stemming indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, *6*(4): 1–33.

Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on learning technologies*, *7*(3): 246–259.

Feldman, R., Sanger, J., *et al.* (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Mattjik, A. A. & Sumertajaya, I. (2011). *Sidik peubah ganda dengan menggunakan SAS*. Bogor (ID): IPB Press.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5): 513–523.

Slamet, C., Rahman, A., Ramdhani, M. A., & Darmalaksana, W. (2016). Clustering the verses of the holy qur'an using k-means algorithm. *Asian Journal of Information Technology*, *15*(24): 5159–5162.

Tala, F. (2003). *A study of stemming effects on information retrieval in Bahasa Indonesia*. Amsterdam (NL): Institute for Logic, Language and Computation, Universiteit van Amsterdam.