Jurnal Muara Sains, Teknologi, Kedokteran, dan Ilmu Kesehatan ISSN 2579-6402 (Versi Cetak)
Vol. 2, No. 1, April 2018: hlm 129-138 ISSN-L 2579-6410 (Versi Elektronik)

PENGGALIAN TEKS DENGAN MODEL *BAG OF WORDS* TERHADAP DATA TWITTER

Wahyuningdiah Trisari Harsanti Putri[1], Retno Hendrowati[2]

[1]Program Studi Teknik Informatika, Universitas Paramadina Jakarta

Email:wahyuningdiah.trisari@paramadina.ac.id

[2] Program Studi Teknik Informatika, Universitas Paramadina Jakarta

Email: retno.hendrowati@paramadina.ac.id

ABSTRAK

Ketersediaan data pada beberapa sosial media memungkinkan penelusuran mengenai hal-hal yang berkaitan terhadap suatu topik atau seorang individu. Penggalian teks (text mining) merupakan proses eksaminasi sumber tertulis dalam jumlah besar guna menghasilkan informasi baru dan untuk mengubah teks yang tidak terstruktur menjadi data yang terstruktur untuk keperluan analisis. Penggalian teks mengidentifikasi fakta-fakta, hubungan-hubungan serta pernyataan yang sekiranya tidak akan ditemukan diantara data teks yang besar. Penelitian ini merupakan eksperimen penggalian teks dengan menggunakan data twit dengan kata kunci nama tiga orang kandidat calon gubernur DKI Jakarta dari twitter stream untuk kurun waktu akhir November sampai dengan Desember 2016. Menggunakan metode penelitian kualitatif dengan tahapan pengambilan data tweet, prapemrosesan teks yang dilanjutkan dengan analisis. Data diambil dan diolah menggunakan API twitter dan Bahasa pemrograman R. Penelitian belum menemukan informasi baru dari proses analisis yang dilakukan. Sepuluh frekuensi kata yang ditemukan untuk tiap dataset, antara lain: agus, kader, dukung, krn,madrid, pks, pilkada, kalah, data, mengejutkan, potensial, aksi, agama, islam, menista, jakarta, politik, program, ahok, pidato, survei, dki, anies, elektabilitas, dan warga.

Kata kunci*:* bag of words, natural language processing, text mining, twitter, R

1. PENDAHULUAN

Latar Belakang

Perkembangan Teknologi Informasi dan Komunikasi (TIK) merupakan salah satu pendorong ketersediaan data yang besar. Pada akhir tahun 2016 data yang beredar di dunia sebesar 1 *zettabytes* dan diperkirakan akan mencapai 30,6 *exabytes* pada tahun 2020 (T. M. Holland, 2016). Ketersediaan data yang diolah menjadi informasi akan membantu individu dan organisasi untuk mengambil keputusan secara lebih tepat. Ledakan data dipicu salah satunya oleh lahirnya media sosial, karena memiliki sifat dua arah atau lebih, seorang pengguna data juga menjadi penghasil data (Obar & Wildman, 2015). Munculnya media sosial telah memberikan pengguna web tempat untuk mengekspresikan dan berbagi pikiran dan pendapat mereka tentang segala macam topik dan peristiwa.

Media sosial didefinisikan sebagai aplikasi berbasiskan internet yang dibangun diatas pondasi teknologi Web versi 2.0 yang memungkinkan penciptaan konten oleh penggunanya (Kaplan & Haenlein, 2010). Sebagai wadah agar orang-orang dapat berkomunikasi, media sosial sangat bermanfaat dan memegang peran cukup penting dalam berkomunikasi. Selain itu, pengguna dapat berbagi informasi berupa kejadian, berbagi informasi, berbagi foto, musik, film dan media mencari pertemanan. Kegunaan media sosial dapat disesuaikan dengan kebutuhan penggunanya dan menggunakan etika dan tata cara yang benar dalam berkomunikasi di media sosial. Beberapa contoh media sosial antara lain Facebook, Twitter, Instagram, Path dan lain sebagainya.

Twitter adalah layanan *microbogging*. Tiap *user* yang telah mendaftarkan sebuah akun dapat memposting pesan. Sebuah pesan twit terdiri dari maksimum 280 karakter. Untuk menautkan pesan dengan pengguna lain maka digunakan simbol "@". Pesan yang telah ditautkan akan muncul pada profil milik pengguna lain yang dituju dan mendorong terciptanya sebuah percakapan (Strickland & Chandler, n.d.). Untuk bergabung dengan percakapan dengan banyak pengguna lainnya, maka digunakan simbol "#" yang disebut *hashtag* yang membuat sebuat kata dapat dibaca oleh mesin. Simbol ini memiliki beberapa tujuan, seperti memberi label, menyimpulkan, serta indikator sebuat topik. Dengan memposting sebuah hashtag dan mencari sebuah hashtag maka dua atau lebih pengguna dapat berpartisipasi pada percakapan yang lebih luas. Seorang pengguna dapat 'mengikuti' pengguna lain, secara otomatis mendapat update dari pesan-pesan mereka dan perubahan-perubahan pada profil mereka.

Penelitian ini bertujuan untuk menemukan fakta-fakta atau hubungan-hubungan yang terdapat di dalam data twit terhadap kata kunci tiga orang kandidat Gubernur DKI. Keberlimpahan data saat kampanye putaran pertama memungkinkan proses ekstraksi data secara besar dalam kurun waktu yang singkat. Pengambilan data tweet dilakukan pada kurun waktu 24 November 2016 sampai dengan 04 Desember 2016. Total data yang berhasil dikumpulkan sejumlah 143.146. Total data ini milik ketiga pasangan calon kandidat. Untuk kemudahan penamaan, hasil kueri nama calon kandidat pertama milik Agus Harimurti Yudhoyono diberi label sebagai dataset I, Basuki Tjahaya Purnama sebagai dataset II serta Anies Baswedan sebagai dataset III sesuai dengan nomor urut pasangan calon pada kampanye putaran pertama.

Tabel 1. Jumlah Data Penelitian

	Dataset I	**Dataset II**	**Dataset III**	**Data Total**
Data twit ketiga calon Gubernur DKI Jakarta	47.612	54.992	40.542	143.146

Tinjauan Pustaka

Pada bagian ini, peneliti melakukan studi literatur terhadap beberapa jurnal penelitian untuk mendukung pemahaman terhadap topik penggalian teks dan model *Bag of Words*.

Natural language processing (NLP)

Natural Language Processing adalah subbidang kecerdasan buatan dan linguistik yang bertujuan agar komputer mengerti bahasa-bahasa manusia yang berbentuk pernyataan ataupun tulisan (Chopra, Prashar, & Chandresh, 2013). Masih merujuk pada tulisan Chopra, terdapat lima fase pada NLP, yaitu:

- Analisis morfologis dan leksikal.
  Leksikon adalah koleksi leksem dalam suatu bahasa yang juga mencakup kata dan ekspresi. Morfologi merupakan pengetahuan tentang bentuk dimana terdapat proses analisis, identifikasi dan deskripsi terhadap struktur kata-kata.
- Analisis sintaksis.

Merupakan ilmu mengenai prinsip dan peraturan untuk membuat kalimat dalam bahasa alami. Termasuk di dalamnya analisis terhadap kata-kata dalam kalimat untuk menggambarkan struktur tata bahasa sebuah kalimat.

- Analisis semantik.
  Semantik berarti ilmu tentang makna kata dan kalimat; pengetahuan mengenai seluk-beluk dan pergeseran arti kata (KBBI, 2016). Analisis semantik melakukan proses abstraksi arti kata dari sebuah konteks. Terdapat pemetaan antara struktur sintaksis dengan objek pada task domain. Contoh kalimat "sepatu biru tak berwarna" akan ditolak oleh *analyser* karena kata "biru" dan "tak berwarna" tidak masuk akal untuk dipadankan bersama.
- Integrasi tulisan.
  Arti sebuah kalimat tunggal bergantung pada kalimat-kalimat yang mendahuluinya dan juga menjadi referensi kalimat-kalimat yang mengikutinya. Contoh kata "nya" pada kalimat "Dia menyukainya" bergantung pada konteks tulisan sebelumnya.
- Analisis pragmatis.
  Ditulis oleh Chopra sebagai mengabstraksi atau memperoleh kegunaan bahasa sebagai tujuan, dimana dibutuhkan pengetahuan terhadap hal yang berlaku alih-alih sekedar makna kata. Contoh: "tutup jendela itu?" dimaknai sebagai permintaan bukan sebagai perintah.

Penggalian teks

Penggalian teks merupakan teknologi yang berusaha untuk mengektraksi informasi yang berguna dari data tekstual yang tidak terstruktur. Penggalian teks merupakan perluasan dari penggalian data (data mining) terhadap data tekstual (He, Zha, & Li, 2013). Informasi ini didapatkan dengan memformulasikan dan mengembangkan pola-pola dan tren melalui pembelajaran pola statistik. Penggalian teks mencakup proses strukturisasi teks input (parsing, dengan tambahan beberapa fitur yang timbul dari sisi linguistik dan pengurangan beberapa fitur lainnya), mendapatkan pola dari data yang telah terstruktur, dan evaluasi dan interpretasi dari keluaran. Informasi dapat dikatakan berkualitas apabila mengandung kombinasi dari relevansi, dan kebaruan. Langkah penggalian teks yang umum dilakukan mencakup kategorisasi teks, pengelompokan teks (clustering), ekstraksi konsep atau entitas, analisis sentimen, peringkasan dokumen dan pemodelan entitas dan relasi.

Model *Bag-of-words*

Model *bag-of-words* (BoW) merupakan representasi sederhana yang digunakan pada natural language processing (NLP) dan information retrieval (IR), juga dikenal sebagai model vector space (McTear, Callejas, & Griol, 2016). Dalam model ini, sebuah teks yang berupa kalimat ataupun dokumen diwakili sebagai kantung (*bag*) multiset dari kata-kata yang terkandung di dalamnya, tanpa memandang urutan kata dan tata bahasa namun tetap mempertahankan keberagamannya. Definisi lain untuk BoW adalah sebuah model yang mempelajari sebuah kosakata dari seluruh dokumen, lalu memodelkan tiap dokumen dengan menghitung jumlah kemunculan setiap kata (Deepu, Pethuru, & Rajaraajeswari, 2016). Berikut contoh teks yang direpresentasikan sebagai BoW

```
(1) Aku merasa sehat hari ini
(2) Aku merasa demam hari ini
(3) Aku berharap aku dapat bermain di luar
```

Berdasarkan kedua kalimat di atas, maka sebuah list dibentuk untuk tiap dokumen sebagai berikut:

```
"aku", "merasa", "sehat", "hari", "ini"
"aku", "merasa",  "demam", "hari", "ini"
"aku", "berharap", "aku",  "dapat", "bermain", "di", "luar"
```

Lalu untuk setiap kata dihitung frekuensinya dan dipetakan kembali ke dalam dokumen

Tabel 2. Frekuensi Kata Dalam Dokumen

	aku	**merasa**	**sehat**	**hari**	**ini**	**demam**	**berharap**	**dapat**	**bermain**	**di**	**luar**
(1)	1	1	1	1	1	0	0	0	0	0	0
(2)	1	1	0	1	1	1	0	0	0	0	0
(3)	2	0	0	0	0	0	1	1	1	1	1

Tabel di atas menggambarkan fitur *training* yang mengandung frekuensi kata dari tiap kalimat dalam sebuah dokumen. Dengan pendekatan ini, tata bahasa dan urutan kata tidak dipentingkan, hanya jumlah kemunculan kata yang diperhatikan.

2. METODE PENELITIAN

Penelitian ini menggunakan beberapa metode penggalian teks yang dirangkum ke dalam sebuah penelitian kualitatif. Tahapan yang dilaksanakan adalah sebagai berikut:

Gambar 1. Tahapan Penelitian

Langkah pertama: Pengambilan data

Sebelum melakukan pengambilan data tweet, maka harus dilakukan pengaturan untuk mendapatkan API Keys dan Tokens dari Twitter. Proses ini dilakukan dengan mendaftarkan aplikasi yang dibangun pada halaman *new application* milik Twitter. Dilanjutkan dengan pengisian data pendukung yang dibutuhkan untuk menghubungkan antara IDE pemograman yang digunakan dalam penelitian dengan twitter. Setelah selesai proses pendaftaran aplikasi, akses dari akun twitter untuk menggunakan aplikasi ini dengan mengklik tombol *Create my Access Token*. Terakhir, salin kode *Consumer key (API key), Consumer Secret, Access Token dan Access Token Secret* ke dalam aplikasi yang hendak digunakan.

Kueri yang digunakan adalah kata “Agus Yudhoyono”, “AHY”, “@AgusYudhoyono”, “Ahok”, “@basuki_btp”, dan “Anies”, “Anies Baswedan”, “@anies_baswedan”. Besaran atau jumlah data yang didapatkan selama periode tersebut adalah 47.612 untuk pasangan calon nomor urut satu, sebanyak 40.542 data untuk pasangan calon nomor urut dua, dan 54.992 untuk pasangan

calon nomor urut tiga yang selanjutnya akan dirujuk dalam penelitian ini sebagai dataset I, dataset II dan dataset III.

Langkah Kedua: Prapemrosesan (*Preprocessing*)

Setelah dokumen diunduh menggunakan bantuan bahasa pemrograman R, maka tweet untuk tiap kandidat dijadikan satu dalam sebuah dokumen *comma separated values* (.csv). Berikut langkah-langkah yang dilakukan:

- *Load* data teks menjadi data frame.
- Membuat *Vcorpus object* dari data frame untuk keperluan langkah pembersihan data.

Setelah itu mulai dilakukan proses pembersihan awal dengan melakukan langkah berikut:

- Menghapus *Uniform Resource Locator* (URL).
- Menghapus teks dengan tanda @mention.
- Menghapus teks dengan tanda #Hashtag.
- Menghapus semua angka yang terdapat dalam teks.
- Menghapus semua tanda baca dalam teks.
- Menghapus semua *white space* dalam teks.
- Mengubah teks menjadi huruf kecil.
- Lakukan pembersihan teks selanjutnya, apabila ditemukan teks masih belum bersih.

Langkah selanjutnya adalah melakukan pembuangan *stopwords. Stopwords* adalah kata-kata yang tidak signifikan untuk melakukan pencarian. Contoh kata-kata ini dalah bahasa Indonesia ialah preposisi (di, ke, dari, dia). Untuk membersihkan teks dari *stopwords*, maka penelitian ini menggunakan *library stopwords* bahasa Indonesia yang telah disusun oleh F.Z Tala (Tala, 2003) yang tersedia pada situs Github. Library ini terdiri dari kompilasi 758 *stopwords*. Berikut contoh teks sebelum dan sesudah proses prapemrosesan.

Tabel 3. Prapemrosesan Teks

Proses	**Teks**
NA	"33","TOP! @TimBRAVO_AHY: Alhamdulillah. Agus-Sylvi melaju terus. Elektabilitas meningkat tajam. Glory. #AgusSylviMakinOKE https://t.co/v5x63bmhdo
Pembersihan awal	top alhamdulillah agus sylvi melaju terus elektabilitas meningkat tajam glory
Pembuangan *stopwords*	top alhamdulillah agus sylvi melaju elektabilitas meningkat tajam glory

Langkah ketiga: Analisis teks / Pemrosesan teks

Analisis teks dilakukan setelah data teks bersih dan siap untuk diproses. Berikut langkah yang dilakukan:

- Membuat sebuah *term-document matrix* (TDM)
    - TDM merupakan matriks yang sering digunakan untuk analisis bahasa. TDM

- Matriks jenis ini lebih mudah untuk dianalisis karena jumlah *terms* diprediksi jauh lebih besar dari jumlah dokumen. Dengan pengaturan ini, maka jumlah baris akan lebih banyak dari jumlah kolom yang akan memudahkan pencarian.

- Menemukan *terms* yang paling sering muncul pada teks.

3. HASIL DAN PEMBAHASAN

Pada bagian ini, peneliti memaparkan hasil eksperimen setelah dilakukan proses pengambilan data melalui API twitter, serta proses prepemrosesan. Eksperimen ini dimulai dengan mengimpor data *stopwords* bahasa Indonesia ke dalam *tools* pemrograman. Selanjutnya setiap kata disimpan sebagai vektor yang akan dicocokan dengan dokumen teks twit yang sudah dibersihkan. Terdapat 758 kata dalam *library* ini.

```
> stopWordsIndo <- read.table("stopwords_id2.txt", header = FALSE, sep = "\n")
> View(stopWordsIndo)
> stop_vec_ind = as.vector(stopWordsIndo$V1)
> class(stop_vec_ind)
[1] "character"
> stop_vec_ind
  [1] "ada"              "adalah"           "adanya"           "adapun"
  [5] "agak"             "agaknya"          "agar"             "akan"
  [9] "akankah"          "akhir"            "akhiri"           "akhirnya"
 [13] "aku"              "akulah"           "amat"             "amatlah"
 [17] "anda"             "andalah"          "antar"            "antara"
 [21] "antaranya"        "apa"              "apaan"            "apabila"
 [25] "apakah"           "apalagi"          "apatah"           "artinya"
 [29] "asal"             "asalkan"          "atas"             "atau"
 [33] "ataukah"          "ataupun"          "awal"             "awalnya"
 [37] "bagai"            "bagaikan"         "bagaimana"        "bagaimanakah"
 [41] "bagaimanapun"     "bagi"             "bagian"           "bahkan"
 [45] "bahwa"            "bahwasanya"       "baik"             "bakal"
 [49] "bakalan"          "balik"            "banyak"           "bapak"
 [53] "baru"             "bawah"            "beberapa"         "begini"
```

Gambar 2. *Stopwords* bahasa Indonesia

Proses selanjutnya adalah membersihkan data twit dari stopwords yang dilanjutkan dengan membentuk sebuah matriks *term-document* untuk dilakukan perhitungan terhadap frekuensi kata yang tertera di dalamnya. Setelah proses analisis teks, ditemukan sebanyak 15.327 *non-sparse* terms dalam dataset I, seperti yang terlihat pada gambar 3. Hal ini berarti terdapat 15.327 kata yang muncul setidaknya satu kali dalam dataset ini. Kata-kata yang memiliki frekuensi tertinggi dapat dilihat selanjutnya pada gambar 4. Sepuluh kata tersebut adalah jakarta, politik, program, ahok, pidato, survei, dki, anies, elektabilitas dan warga. Gambar 5 memberikan visualisasi dari data pada gambar 3.

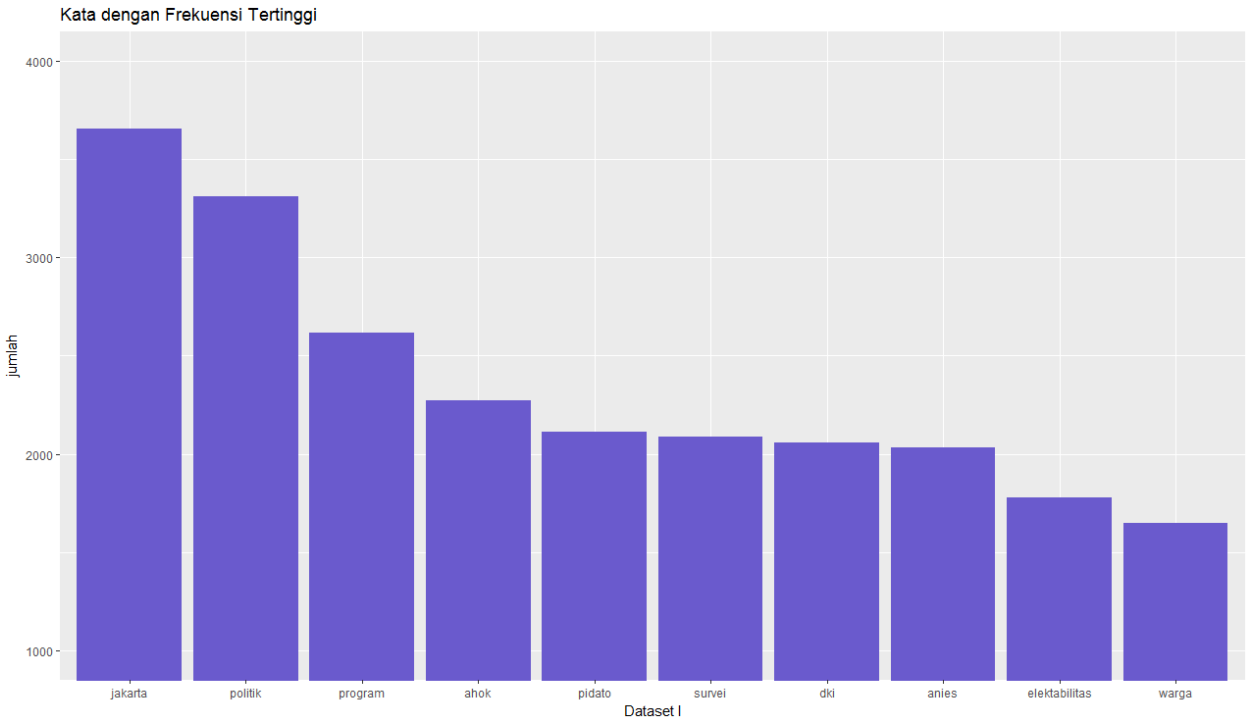
```
> AHYTdm
<<TermDocumentMatrix (terms: 15327, documents: 1)>>
Non-/sparse entries: 15327/0
Sparsity           : 0%
Maximal term length: 240
Weighting          : term frequency (tf)
>
```

Gambar 3. Matriks Dokumen untuk Dataset I

ISSN 2579-6402 (Versi Cetak)
ISSN-L 2579-6410 (Versi Elektronik)

```
> AHY_word_freqs[1:10, ]
                      terms  num
jakarta             jakarta 3653
politik             politik 3309
program             program 2618
ahok                   ahok 2275
pidato               pidato 2113
survei               survei 2087
dki                     dki 2056
anies                 anies 2035
elektabilitas elektabilitas 1780
warga                 warga 1647
>
```

Gambar 4. Sepuluh Besar Frekuensi Kata pada Dataset I dengan Jumlahnya.



Gambar 5. Visualisasi *Term Frequency* Dataset I.

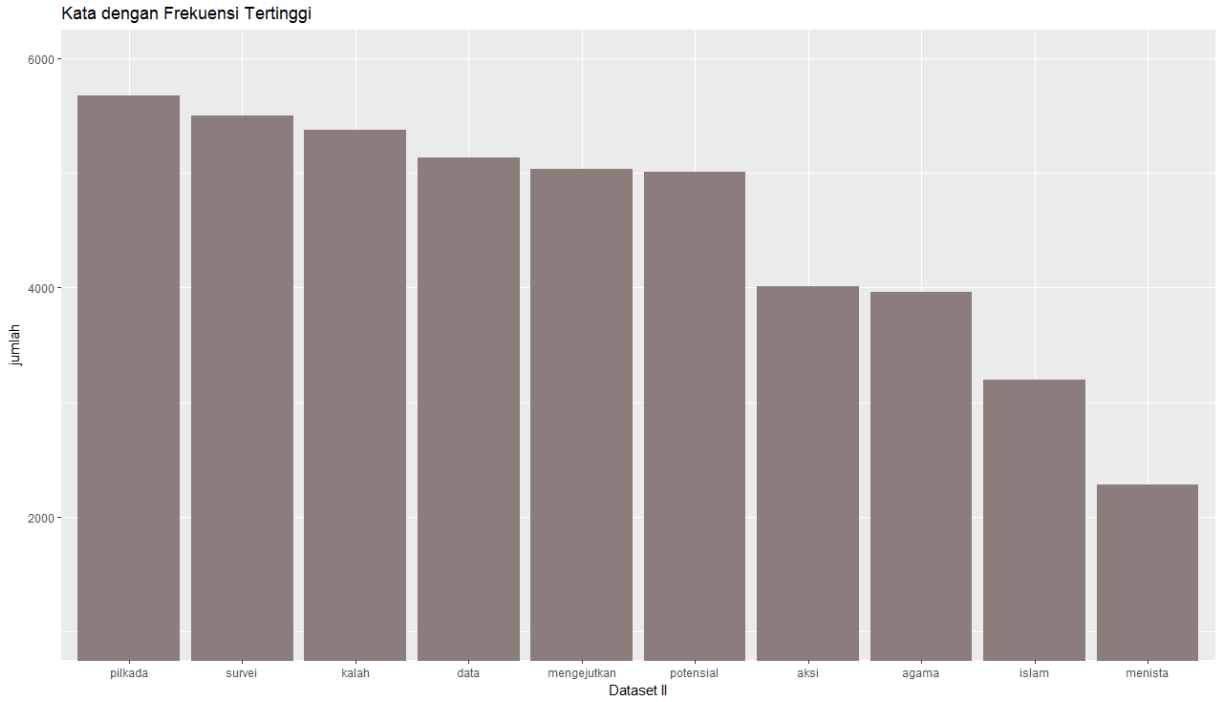
Pada dataset II, terdapat 23.359 *non-sparse terms*, dengan frekuensi sepuluh kata terbanyak yaitu pilkada, survei, kalah, data, mengejutkan, potensial, aksi, agama, islam dan menista.

```
> AhokTdm
<<TermDocumentMatrix (terms: 23359, documents: 1)>>
Non-/sparse entries: 23359/0
Sparsity           : 0%
Maximal term length: 120
Weighting          : term frequency (tf)
>
```

Gambar 6. Matriks Dokumen untuk Dataset II

terms	num
pilkada	5673
survei	5499
kalah	5375
data	5136
mengejutkan	5030
potensial	5007
aksi	4007
agama	3964
islam	3196
menista	2285

Gambar 7. Sepuluh Besar Frekuensi Kata pada Dataset II dengan Jumlahnya.



Gambar 8. Visualisasi *Term Frequency* Dataset II.

Pada dataset III, terdapat 11.850 *non-sparse terms*, dengan frekuensi sepuluh kata terbanyak yaitu ahok, agus, jakarta, dki, kader, warga, dukung, krn, madrid dan pks.

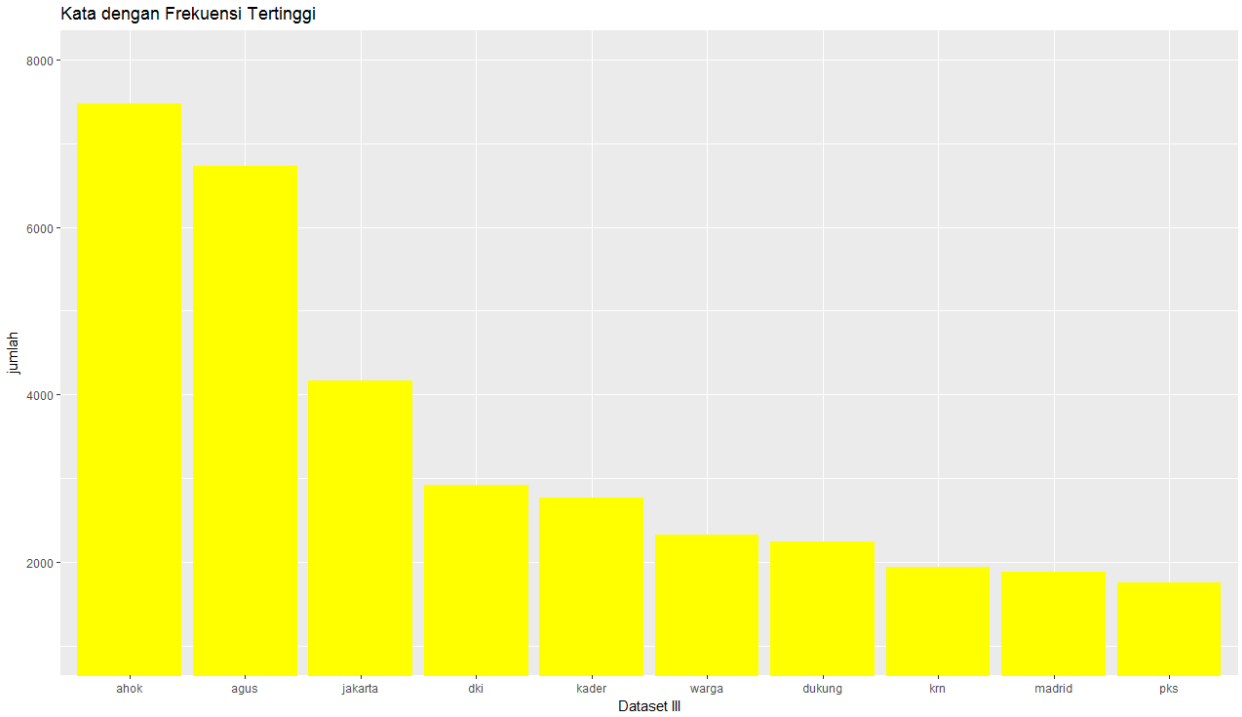
```
> Aniestdm
<<TermDocumentMatrix (terms: 11850, documents: 1)>>
Non-/sparse entries: 11850/0
Sparsity           : 0%
Maximal term length: 276
Weighting          : term frequency (tf)
>
```

Gambar 9. Matriks Dokumen untuk Dataset III

```
> term_frequency[1:10]
   ahok    agus jakarta     dki   kader   warga  dukung     krn  madrid     pks
   7467    6725    4169    2918    2766    2332    2242    1950    1891    1762
>
```

Gambar 10. Sepuluh Besar Frekuensi Kata pada Dataset III dengan Jumlahnya.

ISSN 2579-6402 (Versi Cetak)
ISSN-L 2579-6410 (Versi Elektronik)



Gambar 11. Visualisasi *Term Frequency* Dataset III.

Satu hal yang menarik dari dataset III ini adalah kemunculan kata "madrid" pada frekuensi kata tertinggi. Kata yang tidak berhubungan dengan terminologi politik, sebagaimana terlihat pada kedua dataset lain. Beberapa kata yang sama muncul untuk ketiga dataset adalah "jakarta" dan "dki". Seharusnya kedua kata tersebut masuk ke dalam daftar kata yang dibersihkan karena tidak memberikan wawasan tambahan apapun terhadap analisis teks. Proses pembersihan belum sempurna, karena masih terlihat kata singkatan "krn" muncul pada frekuensi kata tertinggi, hal ini juga menjadi perhatian peneliti karena kata singkatan terdapat cukup banyak pada teks twit.

4. KESIMPULAN DAN SARAN

Setelah dilakukan pengolahan dan analisis terhadap data, sesuai dengan tujuan penelitian, maka dapat disimpulkan bahwa belum ditemukan informasi baru yang terkait dengan hasil pengelompokan dan perhitungan yang disajikan. Model BoW yang digunakan dalam penelitian ini merupakan model sederhana yang cukup efektif untuk menemukan kata-kata (*terms*) yang paling sering dicuitkan oleh warganet terhadap sebuah entitas. Model BoW bisa dikembangkan untuk penelitian lanjutan dibidang analisis sentimen, namun untuk bahasa Indonesia kekurangan satu komponen penting, yaitu sebuah leksikon sentimen bahasa Indonesia yang lengkap dan komprehensif.

Saran untuk penunjang penelitian lebih lanjut ialah pembuatan *library* atau program *stemmer* untuk bahasa Indonesia guna mengembalikan kata-kata ke bentuk dasarnya, mengalihkan kata-kata *slang* atau tidak baku menjadi kata baku untuk memudahkan proses pembersihan data. Pembuatan sebuah leksikon sentimen untuk menunjang penelitian *natural language processing* untuk bahasa Indonesia.

REFERENSI

Chopra, A., Prashar, A., & Chandresh, S. (2013). Natural Language Processing. *International Journal of Technology Enhancements and Emerging Engineering Research*, *1*(4), 131–134.

Deepu, S., Pethuru, R., & Rajaraajeswari, S. (2016). A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction. *International Journal of Advanced Networking & Applications (IJANA)*, 320–323.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*, 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

KBBI. (2016). Entri Semantik. Retrieved January 1, 2018, from https://kbbi.kemdikbud.go.id/entri/semantik

McTear, M., Callejas, Z., & Griol, D. (2016). The Conversational Interface: Talking to Smart Devices. In *The Conversational Interface: Talking to Smart Devices* (pp. 161–185). Springer International Publishing.

Obar, J. A., & Wildman, S. (2015). Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy*, *39*(9), 745–750. https://doi.org/10.1016/j.telpol.2015.07.014

Strickland, J., & Chandler, N. (n.d.). How Twitter Works. Retrieved December 15, 2016, from https://computer.howstuffworks.com/internet/social-networking/networks/twitter2.htm

T. M. Holland. (2016). The World Will Use a Zettabyte of Data in 2016 — How Much Will Your Company Consume? Retrieved November 30, 2016, from https://insights.samsung.com/2016/04/22/the-world-will-use-a-zettabyte-of-data-in-2016-how-much-will-your-company-consume/

Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. *M.Sc. Thesis, Appendix D, pp*, 39–46.