IMPLEMENTASI *TEXT MINING* PADA SOSIAL MEDIA *TWITTER* DALAM MENGANALISIS TOPIK – TOPIK TERKAIT "BARACK OBAMA DAN DONALD TRUMP"

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Jurusan Statistika

Disusun Oleh:

Elisa Fauzia

12 611 093

JURUSAN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2017

IMPLEMENTASI *TEXT MINING* PADA SOSIAL MEDIA *TWITTER* DALAM MENGANALISIS TOPIK – TOPIK TERKAIT "BARACK OBAMA DAN DONALD TRUMP"

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Jurusan Statistika

Disusun Oleh:

Elisa Fauzia

12 611 093

JURUSAN STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2017

HALAMAN PERSETUJUAN PEMBIMBING

Judul	**:**	**Implementasi Text Mining Pada Sosial Media Twitter Dalam Menganalisis Topik-Topik Terkait "Barack Obama" dan "Donald Trump"**
Nama Mahasiswa	**:**	**Elisa Fauzia**
Nomor Mahasiswa	**:**	**12 611 093**

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK

DIUJIKAN

Yogyakarta, 3 Januari 2017

Menyetujui

Dosen Pembimbing

(Prof. Akhmad Fauzy, S.Si., M.Si., Ph.D)

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI TEXT MINING PADA SOSIAL MEDIA TWITTER DALAM MENGANALISIS TOPIK-TOPIK TERKAIT "BARACK OBAMA" DAN "DONALD TRUMP"

Nama Mahasiswa : Elisa Fauzia

Nomor Mahasiswa : 12 611 093

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL 22 FEBRUARI 2017

	Nama Penguji	**Tanda Tangan**
1.	Ir. Ali Parkhan, M.T	.....
2.	Tuti Purwaningsih, S.Stat., M.Si	.....
3.	Prof. Akhmad Fauzy, S.Si., M.Si., Ph.D	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Drs. Allwar, M.Sc., Ph.D

HALAMAN PERSEMBAHAN

Tugas akhir ini penulis persembahkan untuk orang-orang yang disayang:

- ♥ *Ibunda Elisnawati dan Ayahanda Laungke, yang telah memberikan dukungan moril maupun materi serta doa yang tiada henti untuk kesuksesan saya, karena tiada kata seindah lantunan doa dan tiada doa yang paling khusyuk selain doa yang terucap dari orang tua.*
- ♥ *Bapak dan Ibu Dosen pembimbing, penguji, dan pengajar Statistika UII, yang selama ini telah tulus dan ikhlas meluangkan waktunya untuk menuntun dan mengarahkan saya, memberikan bimbingan dan pelajaran yang tiada ternilai harganya, agar saya menjadi lebih baik.*
- ♥ *Kakak Puteri dan Adik-adikku Suci, Dara, Khaira, yang senantiasa memberikan dukungan, semangat, senyum dan doanya untuk keberhasilan saya, cinta kalian telah memberikan kobaran semangat yang menggebu.*
- ♥ *Sahabat dan Teman tersayang, tanpa semangat, dukungan dan bantuan kalian semua saya tidaklah berarti.*

♥Terimakasih yang sebesar-besarnya untuk kalian semua, akhir kata saya persembahkan tugas akhir ini untuk kalian semua, orang-orang yang saya sayangi dan menyayangi saya. Dan semoga tugas akhir ini dapat bermanfaat dan berguna dimasa yang akan datang. Aamiin♥

KATA PENGANTAR

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

Assalaamu'alaikum Wr.Wb.

Segala puji dan syukur penulis panjatkan kehadirat Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya sehingga tugas akhir yang berjudul **Implementasi Text Mining Pada Sosial Media Twitter dalam Menganalisis Topik – Topik Terkait "Barack Obama Dan Donald Trump"**. Shalawat serta salam juga penulis haturkan kepada Nabi Agung Muhammad SAW beserta keluarga, sahabat dan para pengikutnya.

Penulisan tugas akhir ini disusun sebagai salah satu persyaratan yang dipenuhi dalam menyelesaikan jenjang strata satu di Jurusan Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam. Penulis menyadari penyelesaian tugas akhir ini tidak dapat tersusun dengan baik tanpa adanya bantuan dari berbagai pihak oleh karena itu penulis mengucapkan terima kasih yang sebesar-besarnya kepada :

1. Bapak Drs. Allwar, M.Sc.,Ph.D selaku Dekan Fakultas MIPA Universitas Islam Indonesia.
2. Bapak Dr. RB Fajriya Hakim, S.Si., M.Si. selaku Ketua Jurusan Statistika, FMIPA, UII beserta jajarannya.
3. Bapak Prof. Akhmad Fauzy. S.Si., M.Si., Ph.D selaku dosen pembimbing skripsi yang telah memberikan bimbingan, saran dan arahan selama ini.
4. Mama Elisnawati, Bapak Laungke, Kak Putri, Dek Suci, Dek Dara dan Dek Khaira yang selalu memberi dukungan dan doa selama penyusunan tugas akhir ini
5. Syahid Rahman, Fifi Sanjaya, Khoba'drul Eka, Latifa Wulandari, Indra Juniarti, Dewi Ratnasari, Lana Debi, Dyah Kartika, Hanif Rahmat, Shella, dan Maulita yang selalu memberikan semangat dan membantu dalam

penyelesaian tugas akhir ini serta terimakasih atas setiap waktu yang menyenangkan.

6. Teman-teman bimbingan dan teman-teman Statistika UII 2012, terimakasih atas kebersamaannya selama ini.

7. Teman-teman kost Asy-Syfa, terimakasih atas semangat dan doanya.

8. Pihak-pihak yang tidak bisa disebutkan satu persatu, terima kasih atas dukungan dan doa kalian.

Semoga ALLAH SWT membalas kebaikan dan ketulusan semua pihak yang telah membantu dalam penyelesaian skripsi ini. Penulis menyadari dalam penulisan tugas akhir ini masih banyak terdapat kekurangan dan masih jauh dari kata sempurna. Oleh karena itu, segala kritik dan saran sangat dibutuhkan dari semua pihak yang berkepentingan dalam penulisan skripsi ini.

Wassalaamu'alaikum, Wr.Wb.

Yogyakarta, 3 Januari 2017

Penulis

Elisa Fauzia

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pemah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pemah ditulis atau diterbitkan oleh orang iain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daltar pustaka.

Yogyakarta, 3 Januari 2017

Penulis

IMPLEMENTASI *TEXT MINING* PADA SOSIAL MEDIA *TWITTER* DALAM MENGANALISIS TOPIK – TOPIK TERKAIT "BARACK OBAMA DAN DONALD TRUMP"

Oleh : Elisa Fauzia

Program Studi Statistika Fakultas MIPA

Universitas Islam Indonesia

INTISARI

Sebagai salah satu media jejaring sosial yang keberadaannya masih diminati oleh masyarakat luas sampai saat ini, twitter sering digunakan untuk bertukar informasi. Banyaknya kegiatan yang dapat dilakukan menggunakan twitter, maka tidak dipungkiri twitter akan menghasilkan kumpulan data yang besar. Perlu adanya suatu penanganan menggunakan metode khusus untuk mengalisis data pada twitter sehingga tidak terdapat suatu kondisi yang disebut "Rich of Data but Poor of Information". Tugas akhir ini merupakan penerapan metode text mining untuk membahas mengenai perbedaan pemikiran atau topik pembicaraan yang menonjol antara dua orang yang merupakan presiden dan presiden terpilih Amerika Serikat melalui analisis data twitter dari official account mereka. Dimana didapatkan beberapa informasi yang bermanfaat seperti keseringan penggunaan kata-kata yang menyertai kata topik saling berhubungan pada Barack Obama adalah senate. Sedangkan pada Donald Trump adalah will. Diketahui pula tidak terdapat kesamaan (adanya perbedaan) topik yang mereka cuitkan ditwitter. Berdasarkan analisis sentimen Barack Obama lebih banyak berbicara netral dibandingkan dengan Donald Trump karena angka cuitan netral Obama (74.32%) lebih besar dibandingkan dengan angka cuitan netral Trump (50.94%).

Kata Kunci: Presiden, Text Mining, Topik, Twitter, Analisis Sentimen

IMPLEMENTATION TEXT MINING ON TWITTER SOCIAL MEDIA IN ANALYZING TOPICS RELATED TO "BARACK OBAMA AND DONALD TRUMP"

By: Elisa Fauzia

Statistical Studies Program, Faculty Of Mathematics And Natural Sciences

Islamic University Of Indonesia

ABSTRACT

As one of the social networking media whose existence is still in demand by the public until now, Twitter is often used to exchange information. The number of activities that can be carried out using twitter, then there is no doubt Twitter will generate large data sets. The need for a treatment using special methods to mengalisis data on twitter and so there is a condition called "Rich of Data but Poor of Information". The final task is the application of text mining methods to discuss the differences of thought or topic that stands out between two people who are the president and president-elect United States through the analysis of data from the official twitter account of them. Which found some useful information such as the frequency of use of words accompanying the said topic interconnected in Barack Obama is a senate. While Donald Trump is will. Note also there are similarities (their differences) topics they cuitkan ditwitter. Based on the analysis of sentiment Barack Obama talked more neutral compared with Donald Trump as a neutral figure nudge Obama (74.32%) greater than the number of neutral nudge Trump (50.94%).

Keywords: President, Text Mining, Topic, Twitter, Sentiment Analysis

BAB I

PENDAHULAN

1.1. Latar Belakang Masalah

Sebagai makhluk sosial, manusia tidak lepas dari kebutuhan dasar untuk bersosialisasi. Sosialisasi secara umum adalah proses belajar individu untuk mengenal dan menghayati norma-norma serta nilai-nilai sosial sehingga terjadi pembentukan sikap untuk berperilaku sesuai dengan tuntutan atau perilaku masyarakatnya. Salah satu cara bersosialisasi dapat dilakukan melalui komunikasi *verbal* maupun *non verbal* dan secara langsung ataupun tidak langsung. Melalui komunikasi antar individu dapat bertukar kabar atau berita yang menghasilkan suatu informasi (hedisasrawan.blogspot.com, 2013).

Di era modernisasi seperti sekarang ini, sosialisasi antar individu dapat dilakukan dengan komunikasi tidak langsung yaitu melalui media sosial. Media sosial atau sering disebut situs jejaring sosial (*social network sites*) adalah suatu alat (situs media *online*) yang dapat digunakan untuk melakukan komunikasi tanpa adanya interaksi langsung antar individu. Menurut Andreas Kaplan dan Michael Haenlein, mendefinisikan media sosial sebagai "sebuah kelompok aplikasi berbasis internet yang membangun di atas dasar ideologi dan teknologi Web 2.0 dan yang memungkinkan penciptaan dan pertukaran "*user-generated content*" (Anderas, dkk., 2010).

Perkembangan teknologi informasi memungkinkan data dengan jumlah skala yang sangat besar dapat terakumulasi hingga melahirkan gunungan data di bidang ilmu pengetahuan, bisnis dan pemerintahan. Kemampuan teknologi informasi untuk mengumpulkan dan menyimpan berbagai tipe data jauh meninggalkan serta tidak diimbangi dengan kemampuan untuk menganalisis, meringkas, dan mengekstraksi pengetahuan dari data. Pertumbuhan yang pesat dari akumulasi data itu menciptakan kondisi yang disebut sebagai "*rich of data but poor of information*" karena data yang terkumpul tidak dapat memberikan

pengetahuan baru yang dapat digunakan untuk aplikasi yang bermanfaat. Sering dijumpai pada akhirnya gunungan data tersebut dibiarkan begitu saja seperti kuburan data / *data tombs* (Hakim, 2009).

Tumpukan-tumpukan data tersebut salah satu terbesarnya adalah data teks. Data teks setiap detik terus bertambah, baik dari segi *volume* data dan kecepatannya. Pemicu utamanya adalah perkembangan teknologi yang pesat, salah satunya dalam media sosial. Media sosial adalah sebuah media online, dengan para penggunanya bisa dengan mudah berpartisipasi, berbagi, dan menciptakan isi meliputi blog, jejaring sosial, wiki, forum dan dunia virtual. Blog, jejaring sosial, dan wiki merupakan bentuk media sosial yang paling umum digunakan oleh masyarakat di seluruh dunia (ptkomunikasi.wordpress.com, 2012).

Media sosial adalah sebuah media untuk bersosialisasi satu sama lain dan dilakukan secara *online* yang memungkinkan manusia untuk saling berinteraksi. Media sosial meghapus batasan-batasan manusia untuk bersosialisasi, batasan ruang maupun waktu. Media sosial memiliki dampak besar pada kehidupan saat ini. Seseorang yang asalnya "kecil" bisa seketika menjadi besar dengan media sosial, begitupun sebaliknya orang "besar" dalam sedetik bisa menjadi "kecil" dengan media sosial.

Data teks adalah data yang tidak terstruktur di dunia maya dan tidak terkira jumlahnya, bahkan setiap hari terus bertambah. Keberadaan sosial media turut berperan terhadap banjir data yang ada.

Data Mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual, dimana data mining memiliki fungsi umum untuk membentuk *assosiation, sequence, clustering, classification, regretion, forecasting,* dan *solution*. Dalam *data mining* terdapat beberapa metode yang dapat digunakan untuk melakukan analisis data, salah satunya adalah *Text Mining*. *Text Mining* didefinisikan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang

merupakan komponen-komponen dalam data mining dimana salah satu fungsinya adalah kategorisasi (Sanjaya, dkk., 2015).

Text mining adalah salah satu solusi untuk masalah tersebut. Penambangan teks (bahasa Inggris: *text mining*) adalah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data teks, seperti dokumen *Word*, *PDF*, kutipan teks, berita *online*, komentar dan lain-lain. Proses yang umum dilakukan oleh penambangan teks di antaranya adalah perangkuman otomatis, kategorisasi dokumen, penggugusan teks, dan lain-lain (Turban, dkk., 2011).

Terdapat banyak jenis media sosial yang berkembang sampai saat ini, salah satunya adalah situs jejaring sosial (*social network sites*) *Twitter*. *Twitter* merupakan situs jejaring sosial yang keberadaannya masih diminiati oleh masyarkat sampai saat ini. *Twitter* adalah jejaring sosial berupa blog ukuran kecil yang didirikan oleh Jack Dorsey pada bulan Maret 2006. Melalui *Twitter* pengguna dapat mengirim dan membaca pesan, berbagi informasi, menjalin relasi bisnis, menuangkan isi hati dan pikiran dalam bentuk tulisan (sering disebut *tweet*), dengan kapasitas kata yang bisa diunggah dan ditampilkan pada *timeline* pengguna *twitter* mencapai 140 karakter. Sama halnya dengan situs jejaring sosial lain, dalam *Twitter* disediakan suatu mesin pencarian (*search engine*) yang berguna untuk mempermudah pengguna dalam menemukan informasi menggunakan kata kunci. Melalui *search engine* pengguna dapat menemukan lebih banyak informasi yang dibutuhkan terkait topik yang ingin dicari, yaitu lebih dari satu akun yang ada di *twitter* (Imam, 2015).

Twitter sebagai hasil dari perkembangan teknologi informasi memungkinkan setiap waktu untuk menghasilkan kumpulan data yang banyak, dimana setiap detiknya pada saat kehidupan normal rata-rata jumlah *tweet* yang ada dalam twitter adalah 5.700 tweet (tnmedia.com, 2014).

Hal tersebut tidak berlaku jika suatu waktu terjadi peristiwa-peristiwa tertentu yang menyebabkan peningkatan atau penurunan rata-rata jumlah *tweet* per detiknya (Imam, 2015).

Tak dapat dipungkiri bahwa pengguna *Twitter* saat ini telah merambah ke hampir semua kalangan masyarakat, tak terkecuali dua orang sosok yang terkenal sebagai presiden dan presiden terpilih Amerika Serikat, Barack Obama dan Donald Trump.

Brandwatch merupakan perusahaan yang mengembangkan alat atau perangkat lunak untuk memaksimalkan fungsi media sosial. Perusahaan asal Brighton, Inggris itu pada 10 November 2016 meliris daftar pria paling berpengaruh di Twitter. Dan dalam daftar tersebut, terdapat nama presiden dan presiden terpilih Amerika Serikat yang menempati posisi ke-2 dan ke-6.

Oleh karena itu, penulis melakukan penelitian untuk mengetahui apakah ada perbedaan pemikiran atau topik pembicaraan yang menonjol antara dua orang yang sama-sama merupakan presiden Amerika serikat, yaitu Barack Obama dan Donald Trump melalui analisis data twitter dari official account mereka.

1.2. RUMUSAN MASALAH

Adapun rumusan masalah dalam penelitian ini adalah:

a. Bagaimana topik utama dan kata - kata yang melekat pada “Donald Trump dan Barack Obama”?
b. Bagaimana kelompok dari topik-topik lain yang saling berkaitan ?
c. Apakah ada perbedaan pemikiran atau topik pembicaraan yang menonjol antara Barack Obama dan Donald Trump?

1.3. BATASAN MASALAH

Dalam penelitian ini untuk mempertahankan keutuhan obyek penelitian, maka penulis membatasi permasalahan yang akan diteliti sebagai berikut:

a. Data yang dianalisis dan disajikan adalah data teks dari *twitter* yang di ambil dari permintaan kepada sistem sebanyak 2000 *tweet* pada waktu tertentu.

b. Populasi dari penelitian ini adalah berita/dokumen (teks) pada sosial media *twitter* yang termasuk dalam kategori *microblogging* tentang "Donald Trump" dan "Barack Obama".

1.4. JENIS PENELITIAN

Jenis penelitian ini yaitu penelitian kategori aplikasi dengan metode analisis *text mining* dengan menggunakan *software R*.

1.5. TUJUAN PENELITIAN

Berdasarkan rumusan masalah yang telah diuraikan diatas, maka tujuan dari penelitian ini adalah sebagai berikut:

a. Mendeskripsikan topik utama dan kata - kata yang melekat "Donald Trump" dan "Barack Obama".
b. Mengelompokkan topik-topik lain yang saling berkaitan dengan "Donald Trump" dan "Barack Obama"
c. Mengetahui perbedaan pemikiran atau topik pembicaraan yang menonjol antara "Barack Obama" dan "Donald Trump"

1.6. MANFAAT PENELITIAN

Adapun manfaat dari penelitian ini adalah sebagai berikut:

a. Mengetahui hal-hal yang dapat dilakukan untuk mengatasi data besar yang tidak terstruktur terutama data teks di media sosial *twitter*.
b. Mengetahui analisis yang digunakan untuk *text mining* dari media sosial *twitter* dengan menggunakan *software R*.
c. Mengetahui topik utama dan kata - kata yang melekat pada "Donald Trump" dan "Barack Obama".

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Berbagai penelitian tentang pengangguran pernah dilakukan dengan berbagai metode sesuai kebutuhan peneliti. Berikut adalah beberapa penelitian mengenai batubara yang telah dilakukan, seperti dibawah ini.

Faishol (2011) dalam penelitiannya yang berjudul "Implementasi *text mining* untuk mendukung pencarian topik pada *e-library* menggunakan *mobile device*" dengan tujuan untuk memberi peringkat dokumen tersebut saat dicari karena adanya jumlah koleksi dokumen yang begitu besar yang dimiliki oleh sebuah perpustakaan. Penelitian ini menggunakan metode *text mining* yang mengimplementasikan *algoritma cosine similarity* untuk peringkatan dokumen (*page rank*). Data uji coba diperoleh dari perpustakaan pusat Universitas Islam Negeri Malang yaitu berupa abstraksi tugas akhir. Dari hasil pengujian didapat bahwa dokumen relevan yang diterima oleh pengguna mencapai 100% dan akurasi data relevan terhadap data yang diterima pengguna mencapai rata-rata 78,2%.

Susanto (2015) dalam penelitiannya yang berjudul "Visualisasi data teks twitter berbasis bahasa indonesia menggunakan teknik pengklasteran", dengan tujuan untuk pengambilan keputusan selanjutnya, karena di dalamnya dapat dilihat pola data yang sedang diteliti apakah berkecenderungan positif atau negatif. Penelitian ini menggunakan topik isu Pemilu 2014 sebanyak 57294 tweet. Algoritma pengklasteran yang digunakan adalah *K-Means*, *Cascade K-Means* dan *Self-Organizing Map Kohonen*. Hasil yang didapat menunjukkan bahwa *Cascade K-Means* mampu menghasilkan nilai konvergensi kelompok terkecil *SSE* sebesar 7073 dan *Dunn Index* 0,67 dengan distribusi sentimen positif berjumlah 26332 *tweet*, negatif berjumlah 7912 *tweet*, dan netral berjumlah 23050 *tweet*.

Visualisasi menggunakan grafik dua dimensi dengan *evaluator* analisa komponen utama (PCA) pada variabel korelasi *input* 0,95.

Adiyana (2015) dalam penelitiannya "Implementasi *text mining* pada mesin pencarian *twitter* untuk menganalisis topik – topik terkait "KPK dan Jokowi" dengan tujuan sebagai penerapan metode text mining untuk data tweet terkait topik KPK dan topik Jokowi, dimana didapatkan beberapa informasi yang bermanfaat seperti keseringan penggunaan kata-katamenurut aturan asosiasi yang menyertai kata KPK adalah kata polri dan lapor, serta kata jokowi adalah kata widodo, menghadiri, izin, pintu, satu, investor, urus, presiden, nilai, aktif, bahaya, manuver, menang, mulai, relawan, dan sejumlah.

Widhianingsih, dkk (2016) dalam penelitiannya "aplikasi *text mining* untuk *automasi* klasifikasi artikel dalam majalah online wanita menggunakan *naive bayes classifier* (NBC) dan *artificial neural network* (ANN)", dengan tujuan untuk mengkalsifikasikan artikel ke dalam beberapa kategori yang ditentukan. Metode yang digunakan adalah *naive bayes classifier* (NBC) dan *artificial neural network* (ANN). Sebagai perbandingan metode non parametrik tersebut, dilakukan pula analisis menggunakan regresi logistik multinomial. Tingkat akurasi model NBC adalah sebesar 80,71%, model ANN adalah sebesar 75%, dan Reresi Logistik Multinomial adalah sebesar 57,86%. Dengan demikian, dapat dinyatakan bahwa NBC memiliki performansi yang paling baik untuk proses klasifikasi artikel wanita.

Retnawiyati (2015) dalam penelitiannya "Analisis sentimen pada data *twitter* dengan menggunakan *text mining* terhadap suatu produk", penelitan ini melakukan analisis sentimen data dengan mengklasifikasi data *twitter* berbahasa Indonesia pada suatu produk. Data tersebut akan diproses dengan *text mining* untuk menghindari data yang kurang sempurna kemudian data *tweet* diklasifikasi menjadi klasifikasi positif, negatif, dan netral. Klasifikasi ini menggunakan algoritma *naive bayes classifier*.

Abimanyu (2012) dalam penelitiannya "Analisa media sosial *twitter* dengan perhitungan *graph edit distance* untuk mendeteksi rumor pada *trending*

topic siak-ng" dengan tujuan mengidentifikasi rumor pada media sosial *online*. Dengan menggunakan metode kalkulasi *graph edit distance* didapatkan sembilan padanan kata antara *parent node* dan *child node* serta tiga kategori *edge label* dengan kesimpulan ditemukan bahwa rumor sistem siak-ng sedang mengalami *load* tinggi merupakan rumor yang nilai kebenarannya tinggi.

Kurniawan, dkk (2012) dalam penelitiannya "Klasifikasi konten berita dengan metode *text mining*". Dalam penelitian ini data yang digunakan berupa berita yang berasal dari beberapa media online. Berita terdiri dari 4 kategori yaitu politik, ekonomi, olahraga, entertainment. Setiap kategori tediri dari 100 berita; 90 berita digunakan untuk proses training dan 10 berita digunakan untuk proses testing. Hasil dari penelitian ini menghasilkan sistem klasifikasi berita berbasis *web* dengan menggunakan bahasa pemrograman PHP dan *database* MySQL menunjukkan bahwa berita testing bisa terklasifikasi secara otomatis seluruhnya.

3.2. Profil Barack Obama

Barack Hussein Obama II adalah Presiden Amerika Serikat ke-44 yang saat ini sedang menjabat. Ia merupakan orang Afrika Amerika pertama yang menempati jabatan tersebut.

Obama lahir pada 4 Agustus 1961 di Honolulu, Hawaii. Obama merupakan lulusan Universitas Columbia dan Harvard Law School, tempat ia menjadi presiden Harvard Law Review. Ia dulunya seorang penggerak masyarakat di Chicago sebelum mendapat gelar hukumnya. Ia bekerja sebagai jaksa hak-hak sipil di Chicago dan mengajar hukum konstitusi di University of Chicago Law School sejak 1992 sampai 2004. Ia tiga kali mewakili Distrik ke-13 di Senat Illinois mulai tahun 1997 hingga 2004, namun tidak lolos ke tahap Dewan Perwakilan Rakyat Amerika Serikat tahun 2000.

Pada tahun 2004, Obama mendapat perhatian nasional saat berkampanye mewakili Illionis di Senat Amerika Serikat melalui kemenangannya pada pemilu pendahuluan Partai Demokrat bulan Maret, pidatonya di Konvensi Nasional

Demokrat bulan Juli, dan pemilihannya sebagai Senat pada bulan November. Ia memulai kampanye presidennya tahun 2007, dan pada tahun 2008, setelah kampanye pendahuluan melawan Hillary Rodham Clinton, Obama memenangkan mayoritas suara delegasi dalam pemilu pendahuluan partai Demokrat untuk dijadikan calon presiden. Ia kemudian mengalahkan calon dari Partai Republik John McCain dalam pemilihan umum presiden tahun 2008, dan dilantik sebagai presiden pada tanggal 20 Januari 2009. Sembilan bulan kemudian, Obama dinyatakan sebagai pemenang Hadiah Nobel Perdamaian 2009. Ia terpilih lagi sebagai presiden pada November 2012, mengalahkan Mitt Romney dari Partai Republik, dan dilantik untuk kedua kalinya pada tanggal 20 Januari 2013.

Pada masa jabatan pertamanya, Obama mengesahkan undang-undang stimulus ekonomi sebagai tanggapan terhadap resesi 2007–2009 di Amerika Serikat dalam bentuk American Recovery and Reinvestment Act of 2009 dan Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010. Inisiatif besar dalam negeri lainnya pada masa pemerintahannya adalah Patient Protection and Affordable Care Act; Dodd–Frank Wall Street Reform and Consumer Protection Act; Don't Ask, Don't Tell Repeal Act of 2010; Budget Control Act of 2011; dan American Taxpayer Relief Act of 2012. Di bidang kebijakan luar negeri, Obama mengakhiri keterlibatan militer A.S. dalam Perang Irak, menambah jumlah tentara di Afganistan, menandatangani perjanjian pengendalian senjata New START bersama Rusia, memerintahkan intervensi militer A.S. di Libya, dan melaksanakan operasi militer yang berujung pada kematian Osama bin Laden. Pada bulan mei 2012, ia menjadi presiden A.S. pertama yang mendukung pengesahan pernikahan sesama jenis secara terbuka.

3.3. Profil Donald Trump

Donald John Trump adalah pebisnis, tokoh televisi realita, politikus, dan Presiden terpilih Amerika Serikat. Sejak 1971, ia memimpin The Trump Organization, perusahaan induk utama untuk semua usaha properti dan kepentingan bisnis lain miliknya. Sepanjang karier bisnisnya, Trump telah membangun gedung perkantoran, hotel, kasino, lapangan golf, dan fasilitas

bermerek lainnya di seluruh dunia. Ia terpilih sebagai presiden Amerika Serikat ke-45 pada pilpres 2016 dari Partai Republik; ia mengalahkan calon dari Partai Demokrat, Hillary Clinton. Ia akan dilantik pada tanggal 20 Januari 2017.

Trump lahir dan besar di New York City pada 14 Juni 1946. Ia mendapat gelar sarjana dari jurusan ekonomi Wharton School di Universitas Pennsylvania pada tahun 1968. Tahun 1971, ia mengambil alih operasi perusahaan properti dan konstruksi milik bapaknya, Fred Trump. Trump tampil di berbagai ajang Miss USA yang penyelenggaraannya dikuasai Trump sejak tahun 1996 sampai 2015. Ia juga tampil secara mendadak di sejumlah film dan seri televisi. Ia sempat mencalonkan diri sebagai presiden dari Partai Reformasi pada tahun 2000, namun mengundurkan diri sebelum pemungutan suara dimulai. Ia merupakan pembawa acara dan produser The Apprentice, seri televisi realita di NBC, pada tahun 2004 sampai 2015. Pada 2016, ia terdaftar di Forbes sebagai orang terkaya ke-324 di dunia dan ke-113 di Amerika Serikat dengan kekayaan bersih $4,5 miliar.

Pada Juni 2015, Trump mengumumkan pencalonan dirinya sebagai presiden dari Partai Republik dan langsung menjadi calon unggulan. Bulan Mei 2016, para pesaingnya di Partai Republik menghentikan kampanyenya masing-masing. Bulan Juli 2016, ia secara resmi dicalonkan sebagai presiden pada Konvensi Nasional Republik 2016. Kampanye Trump mendapat liputan media dan perhatian luas di dalam maupun luar negeri. Banyak pernyataan Trump dalam berbagai wawancara, Twitter, maupun kegiatan kampanyenya yang memicu kontroversi atau terbukti keliru. Sejumlah kegiatan kampanye Trump sepanjang pemilihan pendahuluan dibarengi oleh unjuk rasa. Setelah Trump menang pemilu, ia memulai proses transisi pemerintahan. Pada usia 70 tahun, ia merupakan orang tertua yang menjabat sebagai presiden Amerika Serikat.

Kebijakan Trump meliputi renegosiasi perjanjian dagang A.S.–Cina, penolakan terhadap beberapa perjanjian dagang seperti NAFTA dan Kemitraan Trans-Pasifik, penegakan hukum imigrasi yang lebih ketat serta membangun tembok di sepanjang perbatasan A.S.–Meksiko, reformasi perawatan veteran, pembatalan dan penggantian Undang-Undang Layanan Kesehatan Terjangkau

(Affordable Care Act), dan pemotongan pajak. Setelah serangan Paris November 2015, Trump mengusulkan penghentian sementara imigrasi Muslim ke Amerika Serikat; ia kemudian mengubah rencana kebijakannya menjadi "pemeriksaan latar sangat ketat" dari negara-negara tertentu.

BAB III
LANDASAN TEORI

3.1. Populasi dan Sampel

Keselurhan objek pengamata yang menjadi perhatian peneliti baik tak hingga maupun terhingga disebut populasi. Semua anggota yang ada dalam populasi disebut anggota populasi dan banyaknya anggota yang ada dalam populasi disebut ukuran populasi. Dalam inferensi statistika tentunya ingin memperoleh kesimpulan mengenai populasi, walaupun tidak mungkin atau tidak praktis untuk mengamati keseluruhan individu yang menyusun populasi. Oleh karena itu terpaksa menggantungkan kepada sebagian anggota populasi untuk membantu peneliti menarik kesimpulan mengenai populasi tersebut. Ini mengarah kepada pengertian sampel.

Sampel adalah suatu himpunan bagian dari populasi. Sampel diharapkan akan mewakili keadaan populasi (repsentatif). Banyaknya anggota dalam sampel disebut ukuran sampel. Keterwakilan populasi dipengaruhi oleh ukuran sampel, cara pengambilan sampel, cara memperoleh data atau mengumpulkan data dan ketelitian (dalam tingkat kekeliruan dan ketidakpastian) kesimpulan yang diinginkan. Oleh karena itu dalam memilih sampel harus mengikuti prosedur tertentu yang dipelajari dalam teknik sampling (Jaka, 2013).

Jenis dan Metode Sampling Sampling secara garis besar dapat dikelompokkan menjadi dua kelompok, yaitu Probability sampling dan Nonprobability sampling. Adapun Probability sampling menurut Sugiyono adalah teknik sampling yang memberikan peluang yang sama bagi setiap unsur (anggota) populasi untuk dipilih menjadi anggota sampel. Sedangkan Nonprobability sampling menurut Sugiyono adalah teknik yang tidak memberi peluang/kesempatan yang sama bagi setiap unsur atau anggota populasi untuk dipilih menjadi sampel (eurapendidikan.com, 2015).

3.2. Statistika Deskriptif dan Statistika Inferensi

Studi tentang statistika telah menjadi sangat populer pada sekitar 3 dekade terakhir. Peran statistika sebagai alat analisis dalam berbagai riset ilmiah semakin meningkat dengan adanya perkembangan kemajuan komputer dengan berbagai program statistiknya. Akibat dari hal tersebut, statistika saat ini digunakan sebagai alat analisis dalam hampir semua bidang profesi, dan kesehatan sampai dengan olahraga, termasuk juga dalam ilmu ekonomi dan bisnis.

Statistics atau ilmu statistik atau statistika adalah sebuah ilmu yang mempelajari teknik-teknik pengumpulan, pengorganisasian (pengaturan), analisis dan interpretasi atas informasi data. Lebih jauh lagi para ahli menggolongkan statistika aplikasi dalam dua golongan besar yaitu statistika deskriptif dan statistika inferensi. Statistika deskriptif berisi metode pengumpulan dan penyajian data kemudian interpretasi dan analisis data populasi tersebut secara langsung. Apabilia dalam analisis deskritif dipelajari sifat data sampel, tekanannya hanya terletak pada cara mencari berbagai nilai besaran-besaran data sampel tersebut (dinamakan statistik, bukan parameter) tetapi belum membahas penggunaan data sampel tersebut untuk proses inferensi. Disisi yang lain, statistika inferensi berkaitan dengan penggunaan data sampel tersebut untuk berbagai tujuan analisis dan interpretasi sifat-sifat populasi, diantaranya penaksiran parameter dan pengujian hipotesis, serta peramalan.

Jika data statistika disajikan apa adanya, akan sulit bagi para pembaca untuk membuat interpretasi atas lapan tersebut. Agar bisa dibaca dan diinterpretasikan dengan mudah dan cepat maka data tersebut harus diolah dengan metode statistika deskriptif. Metode-metode statistika deskriptif berupa pembuatan jenis grafik dan perhitungan atau pencarian ringkasan ukuran-ukuran numerik.

3.3. Analisis *Text Mining*

3.3.1. Perangkat Lunak (software) R

R pertama kali muncul pada tahun 1996, ketika profesor statistik Robert Gentleman dan Ross Ihakadari University of Auckland, New Zealand merilis kode sebagai suatu *free software package*.

R telah banyak digunakan oleh sekitar 1-2 juta pengguna dan didukung dengan baik oleh proyek *software* bebas dan *open source*. R dapat berjalan pada *GNU/linux*, *unix*, *MacOS* dan *Windows*, mempunyai kemampuan grafis yang sangat baik, dapat menangani adanya *missing data*, tersedia lebih dari 4.300 *package* yang mudah digunakan dan mempunyai layanan bantuan *online* dan dokumentasi terkait.

R dapat sangat berguna apabila para ahli statistik, insinyur dan ilmuwan dapat meningkatkan kode software atau atau menulis variasi untuk tugas-tugas tertentu. *Package* yang dibuat untuk *R* menambahkan *algoritma* yang canggih, berwarna dan grafik bertekstur serta teknik pertambangan (*mining*) untuk menggali lebih dalam *database*.

3.3.2. twitteR Package

Paket *twitteR* ini dimaksudkan untuk memberikan akses ke *Twitter API* dalam *R*. Pengguna dapat membuat akses data *Twitter* dalam jumlah besar untuk *data mining* dan tugas-tugas lainnya. Paket ini ditujukan untuk dikombinasikan dengan paket *ROAuth* dikarenakan sejak pada Maret 2013 *Twitter API* memerlukan penggunaan otentikasi *OAuth*.

3.3.3. Sosial Media dan Twitter

"Sosial media" adalah istilah yang luas yang diberikan untuk menggambarkan evolusi terbaru dari internet dan platform komunikasi berbasis web yang memungkinkan pengguna untuk dengan cepat terhubung dan berinteraksi dalam berbagai format yang berbeda. Sebuah situs media sosial adalah sebuah platform yang memungkinkan konten yang dibuat pengguna

muncul melalui interaksi dan kolaborasi dalam komunitas virtual. Ini berbeda dengan situs sebelumnya dan bentuk lain dari media penyiaran di mana pengguna terbatas pada tampilan pasif dari suatu konten.

Sosial media merupakan alat komunikasi yang *powerful* yang memiliki dampak yang signifikan terhadap reputasi organisasi dan profesional. Sosial media didefinisikan sebagai media yang dirancang untuk disebarluaskan melalui interaksi sosial, dibuat dengan menggunakan *highly accessible* dan *scalable publishing techniques online*. Contoh: *LinkedIn*, *Facebook*, *Twitter*, *YouTube*, *Flickr*, *iTunes U*, *Second Life* dan *MySpace*.

Twitter adalah layanan microblogging populer di mana pengguna mengirimkan pesan yang sangat singkat yaitu kurang dari 140 karakter dan rata-rata 11 kata per pesan. Hal ini nyaman untuk penelitian karena adanya jumlah pesan yang sangat besar, kebanyakan tersedia untuk publik, dan secara teknis dapat diperoleh denganlebih sederhana dibandingkan dengan *scraping blog* dari *web*.

Microblogging merupakan fenomena yang relatif baru yang didefinisikan sebagai "suatu bentuk blogging yang memungkinkan pengguna menulis teks update singkat (biasanya kurang dari 200 karakter) tentang kehidupan pengguna di mana saja dan dikirim ke teman-teman melalui pesan teks, *instant messaging* (IM), *email* atau *web*.

3.3.4. Text Mining

Text mining merupakan penerapan konsep dan teknik *data mining* untuk mencari pola dalam teks dan menemukan informasi atau tren terbaru yang sebelumnya tidak terungkap, dengan memproses dan menganalisa data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian teks dengan yang lainnya berdasarkan aturan-aturan tertentu (Kunaifi, 2009). Proses data mining untuk data dokumen atau teks memerlukan lebih banyak tahapan, mengingat data teks memiliki karakteristik yang lebih kompleks daripada data biasa. Selain itu text

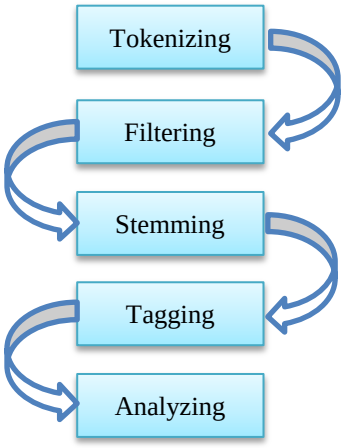
mining juga diartikan sebagai kegiatan menambang data dari data yang berupa teks atau dokumen, dengan tujuan mencari kata-kata yang dapat mewakili apa yang ada dalam dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Ungkapan "*text mining*" umumnya digunakan untuk menunjukkan sistem yang menganalisis *natural language text* dalam jumlah besar dan mendeteksi pola penggunaan bahasa dalam upaya untuk mengekstrak informasi yang mungkin berguna (Sebastiani, 2002).

Menurut Loreta Auvil dan Duane Searsmith dari University of Illinois, karakteristik dokumen teks meliputi:

a. Database teks yang berukuran besar
b. Memiliki dimensi yang tinggi, yakni satu kata merupakan satu dimensi
c. Mengandung kumpulan kata yang saling terkait (frase) dan antara kumpulan kata satu dengan lain dapat memiliki arti yang berbeda
d. Banyak mengandung kata ataupun arti yang bias (ambiguity)
e. Dokumen email merupakan dokumen yang tidak memiliki struktur bahasa yang baku, karena di dalamnya terkadang muncul istilah slank seperti ”r u there?”, ”helllooo bosss, whatzzzzzzz up?”, dan sebagainya.

3.3.4.1.Tahapan dalam text mining

Berdasarkan ketidakteraturan struktur data teks, maka proses *text mining* memerlukan beberapa tahap awal yang pada intinya adalah mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Bentuk perubahan yang dilakukan adalah ke dalam *spreadsheet*, kolom menunjuk dokumen dan baris menunjuk kata, sedangkan selnya menunjuk frekuensi kata dalam dokumen.



Gambar 2.1 Tahapan dalam Text Mining

a. *Tokenizing*

Tahap *tokenizing* adalah tahap pemotongan *string input* berdasarkan tiap kata yang menyusunnya. Tokenisasi secara garis besar memecah sekumpulan karakter dalam suatu teks ke dalam satuan kata dan membedakan karakter-karakter tertentu yang mana saja yang dapat diperlakukan sebagai pemisah kata atau bukan. Namun untuk karakter petik tunggal ('), titik (.), semikolon (;), titik dua (:) atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata. Dalam memperlakukan karakter-karakter dalam teks sangat tergantung sekali pada kontek aplikasi yang dikembangkan. Pekerjaan *tokenisasi* ini akan semakin sulit jika juga harus memperhatikan struktur bahasa (*grammatikal*).

b. *Filtering*

Tahap *filtering* adalah tahap mengambil kata-kata penting dari hasil *token* yang dapat dilakukan dengan menggunakan *algoritma stop list* (membuang kata yang kurang penting) atou *word list* (menyimpan kata penting). Yang termasuk *stoplist* adalah “yang”, “di”, “dari”, dan lain-lain.

c. *Stemming*

Tahap *stemming* adalah tahap mencari akar kata dari setiap kata hasil *filtering*. *Stemming* adalah proses untuk menggabungkan atau memecahkan setiap

varian-varian suatu kata menjadi kata dasar. *Stem* (akar kata) adalah bagian dari akar yang tersisa setelah dihilangkan imbuhannya (awalan dan akhiran).

d. *Tagging*

Tahap *tagging* adalah tahap mencari awal dari tiap kata lampau atau kata hasil *stemming*. Tahap ini tidak digunakan pada teks berbahasa indonesia karena kata dalam bahasa indonesia tidak mempunyai bentuk lampau.

e. *Analyzing*

Proses *analyzing* adalah proses analisa dari hasil proses *tagging* sehingga diketahui seberapa jauh tingkat keterhubungan antar kata-kata dan antar dokumen yang ada.

3.3.5. *Word Cloud*

Word Cloud adalah salah satu hasil dari metode *text mining*, yang menampilkan kata kata populer terkait dengan kata kunci internet dan data teks. Semakin sering kata muncul dalam teks yang dianalisis, semakin besar ukuran kata muncul dalam gambar yang dihasilkan. Biasanya *word cloud* mem*plot*kan frekuensi kata dilihat dari ukuran katanya (*by the size of the word*). *Word Cloud* sering digunakan ntuk menyoroti istilah populer atau tren berdasarkan frekuensi penggunaan kata (PBC, 2013). *Word Cloud* merupakan pendekatan yang dapat menjelaskan pertanyaan penelitian dengan sangat cepat dan mudah, kita dapat menjelajahi *Word Cloud* secara singkat dan dapat melakukan analisis yang komprehensif (Graham, I. Milligan, & S. Weingart).

Yang perlu diperhatikan ketika menggunakan *word cloud* antara lain:

a. Data yang digunakan merupakan data yang bermakna (*meaningful state*).
b. Perlu menggunakan *software* yang dapat bekerja dalam hal ejaan dan tanda baca, harus dipastikan bahwa semua ejaan benar dan terstandardisasi sehingga setiap kata hanya memiliki satu ejaan, selain itu tanda baca juga dapat mempengaruhi hasil.
c. *Word Cloud* sering gagal dalam mengelompokkan kata-kata yang sama, sehingga diperlukan suatu metode yang dapat meminimalisasi dampak ini.

3.3.6. *Text Clustering*

Text clustering membantu dalam pengambilan dengan menciptakan hubungan antara dokumen yang sama, yang pada gilirannya memungkinkan dokumen terkait menjadi diambil setelah salah satu dokumen telah dianggap relevan dengan permintaan (Martin, 1995).

Aplikasi utama dari *clustering* dalam *text mining* adalah:

a. *Simple clustering*. Hal ini mengacu pada pembentukan cluster fitur text. Sebagai contoh: pengelompokan hits yang dihasilkan oleh*search engine*.
b. *Taxonomy generation*. Hal ini mengacu pada generasi kelompok hirarkis. Sebagai contoh: sebuah cluster yang berisi teks tentang produsen mobil merupakan induk dari cluster anak yang berisi teks tentang model-model mobil.
c. *Topic extraction*. Hal ini mengacu pada ekstraksi fitur yang paling khas dari suatu kelompok. Sebagai contoh: karakteristik yang paling khas dari dokumen dalam setiap topik dokumen.

3.3.7. *K-Means*

Algoritma K-Means adalah *algoritma clustering* yang paling popular dan banyak digunakan dalam dunia industri. Metode *k-means* telah terbukti efektif dalam memproduksi hasil pengelompokan yang baik untuk beberapa aplikasi praktis. Namun, algoritma langsung dari metode *k-means* membutuhkan waktu yang sebanding dengan produk dari jumlah pola dan jumlah *cluster* per-*iterasi*. *Algoritma k-means* merupakan komputasi yang sangat mahal terutama untuk dataset besar. Langkah-langkah pada *algoritma K-Means*:

a. Menentukan banyak *cluster* yang akan dibentuk
b. Menentukan koordinat titik tengah (*centroids*) awal pada setiap *cluster*
c. Menentukan jarak setiap obyek terhadap koordinat titik tengah (*centroids*)
d. Mengelompokkan obyek-obyek tersebut berdasarkan pada jarak minimumnya
e. Menghitung ulang nilai *centroids* dengan menghitung nilai *mean* data dari masing-masing *cluster*

f. Melakukan pengluangan langkah 3-5 hingga nilai *centroids* tidak lagi mengalami perubahan.

Beberapa kelebihan *K-Means*, yaitu:

a. Selalu mampu melakukan klasterisasi
b. Tidak membutuhkan operasi matematis yang rumit
c. Beban komputasi realtif lebih ringan sehingga klasterisasi bisa dilakukan dengan cepat walaupun relatif tergantung pada banyak jumlah data dan jumlah *cluster* yang ingin dicapai.

Beberapa hal yang dianggap sebagai kelemahan *K-Means*, yaitu:

a. Adanya keharusan menentukan banyaknya cluster yang akan dibentuk
b. Hanya dapat digunakan dalam data yang *mean*-nya dapat ditentukan
c. Tidak mampu menangani data yang mempunyai penyimpangan-penyimpangan (*noisy data* dan *outlier*).

Sedangkan menurut Berkhin, beberapa kelemahan Algoritma *K-Means* adalah:

a. Sangat bergantung pada pemilihan nilai awal *centroid*
b. Tidak jelas berapa banyak *cluster k* yang terbaik
c. Hanya bekerja pada atribut numerik.

BAB IV

METODE PENELITIAN

Penelitian ini merupakan penelitian kualitatif, yang dilakukan untuk mengetahui kicauan (*tweet*) serta topik-topik yang sering dikicaukan oleh Barack Obama dan Donald Trump. Penelitian ini menggunakan data-data teks pada media sosial *Twitter* yang berkaitan dengan kata kunci "BarackObama" dan "realDonaldTrump". Tujuan akhir penelitian ini adalah dapat mendeskripsikan topik utama dan kata-kata yang melekat serta sering dikicaukan pada "Barack Obama" dan "Donald Trump", serta mengelompokkan topik - topik lain yang saling berkaitan.

4.1. Populasi

Populasi dari penelitian ini adalah berita/dokumen (teks) pada sosial media Twitter yang termasuk dalam kategori microblogging adalah kicauan-kicauan (tweets) dari akun Barack Obama (@BarackObama) dan Donald Trump (@realDonaldTrump).

4.2. Sampel dan Teknik Pengambilan Sampel

Sampel yang diambil adalah data teks dari twitter yang diambil dengan permintaan kepada sistem sebanyak 2000 *tweets* yang diambil berasal dari akun Barack Obama (@BarackObama) dan Donald Trump (@realDonaldTrump). Setiap Tweet memiliki parameter yaitu *status*, *in replay to status id*, *latitude*, *longitude*, *place id*, *display coordinates*, *trims user*, dan *includes entities*. Dalam penelitian ini hanya menggunakan satu parameter, yaitu status sebagai data yang akan di analisis.

4.3. Sumber Data

Data yang digunakan adalah data primer yang dikumpulkan dari *tweets* @BarackObama @realDonaldTrump pada media sosial *twitter* menggunakan *software R* dengan perintah sebagai berikut:

userTimeline('BarackObama',n=2000,maxID=NULL,sinceID=NULL,includeRts=FALSE,excludeReplies=FALSE)

userTimeline('realDOnaldTrump',n=2000,maxID=NULL,sinceID=NULL,includeRts=FALSE,excludeReplies=FALSE)

4.4. Metode Pengumpulan Data

Data di ambil dengan cara men*download* dan mengumpulkan *tweets* pada akun Barack Obama dan Donald Trump dengan menggunakan perintah *"userTimeline*", yang kemudian menjadi suatu kumpulan dari banyak "keranjang" data teks. Pengolahan data penelitian menggunakan bantuan software open source R dengan tambahan paket "twitteR", "tm", "wordcloud", dan "RColorBrewer".

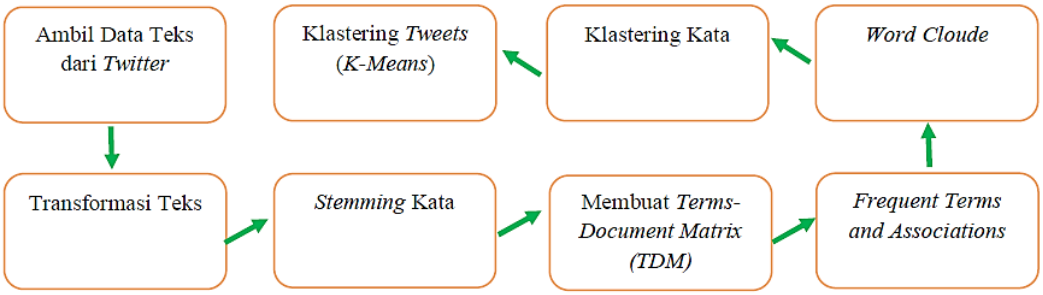
4.5. Tahapan Analisis Data

Secara umum, tahapan melakukan analisis text mining dapat digambarkan dalam diagram alir berikut.

Gambar 4.1. *Tahapan Analisis Data*

4.6. Pengolahan Data

Pengolahan data teks yang didapatkan dari *tweet twitter* diolah menggunakan Sofware R dengan tahapan ditunjukkan pada diagram alir berikut:



Gambar 4.2. *Diagram Alir Pengolahan Data*

BAB V

PEMBAHASAN

Pada analisis *text mining* dengan menggunakan akun twitter @BarackObama dan @realDonaldTrump dengan target data *tweet* yang digunakan sebanyak 2000 *tweet* di masing-masing akun. Tetapi, karena keterbatasan *bandwitch* dalam analisis ini hanya mampu mendapatkan *tweet* sebanyak 388 *tweet* untuk akun @BarackObama dan 817 *tweet* untuk akun @realDonaldTrump.

Berikut adalah tahapan dan hasil analisis *text mining*:

5.1. Tahapan Analisis

Dalam menyelesaikan penelitian ini ada beberapa tahapan yang dilakukan, mulai dari awal mengumpulkan data hingga data tersebut dianalisis. Untuk memperjelas tahapan analisis *text mining* akan dijelaskan alur penelitiannya sebagai berikut:

a. *Data Readings* (Pembacaan Data)

   Mengumpulkan semua dokumen yang terkait dengan konteks yang diinginkan yang kemudian disimpan kedalam sebuah dokumen dengan nama *myCorpus*.

b. Ekstraksi Fitur

   Ekstraksi fitur sendiri ada empat tahap, yaitu *tokenizing*, *filtering*, *stemming*, dan *tagging*. Dimana *tokenizing* merupakan pemotongan *string* berdasarkan kata yang menyusunnya (contoh: jika ada sebuah kalimat "We asked. You answered" maka apabila melewati tahapan *tokenizing* akan menjadi "*we*, *asked*, *you*, *answered*". Selanjutnya adalah *filtering*, dalam tahapan ini kata-kata yang tidak penting akan dibuang (contoh: *we*, *you*, *is*, *and*). Dalam tahapan *stemming*, varian-varian kata akan dipecah menjadi suatu kata dasar (biasanya dihilangkan imbuhan awalan dan akhirannya).

Tahapan *tagging* adalah tahapan mencari awal dari tiap kata lampau dari hasil *stemming* (contoh: *bought* → *buy*).

c. Pembobotan Term

Dokumen-dokumen (corpus) yang sudah diorganisir digunakan untuk membuat 'term-document matrix' atau TDM (matrik yang berisi berbagai 'istilah-dan-dokumen' nya). Dalam pembobotan term akan diketahui berapa *terms* dan *document* yang dianalisis, serta maksimal panjang *term* dalam sebuah dokumen juga akan diketahui.

d. Analisis

Dalam proses analisis dari hasil ektraksi fitur akan diketahui seberapa jauh tingkat keterhubungan antar kata-kata dan dokumen yang ada.

5.2. Analisis

Berikut ini adalah keluaran-keluaran yang dihasilkan dari analisis *text mining* berbasis media sosial *Twitter* dengan kata kunci "Barack Obama" dan "RealDonaldTrump".

5.2.1. Topik Utama dan Kata yang Melekat

Dengan menggunakan *package* (ggplot2) dan *syntax* sebagai berikut:

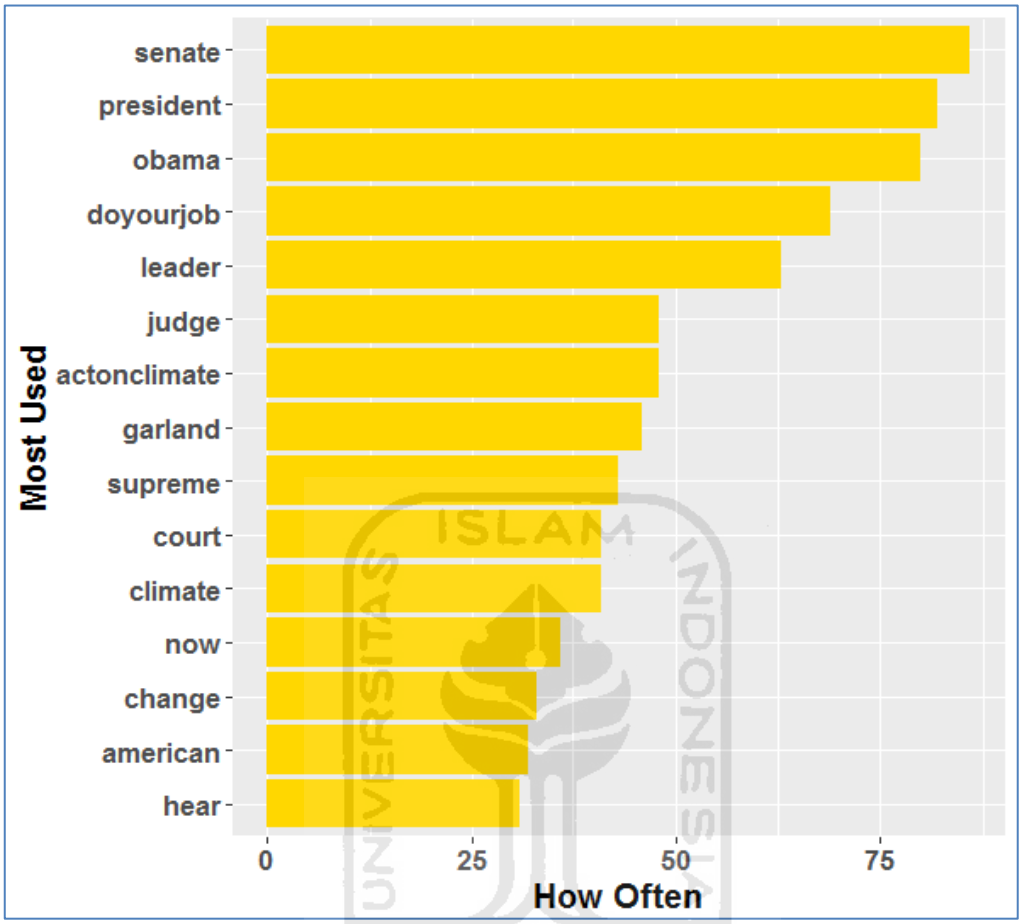
Akun @BarackObama:

- obama_plot <- ggplot(df, aes(x = reorder(term, freq), y = freq)) + geom_bar(stat = "identity", fill = "gold") + xlab("Most Used") + ylab("How Often") + coord_flip() + theme(text=element_text(size=15,face="bold"))
- obama_plot

Akun @RealDonaldTrump:

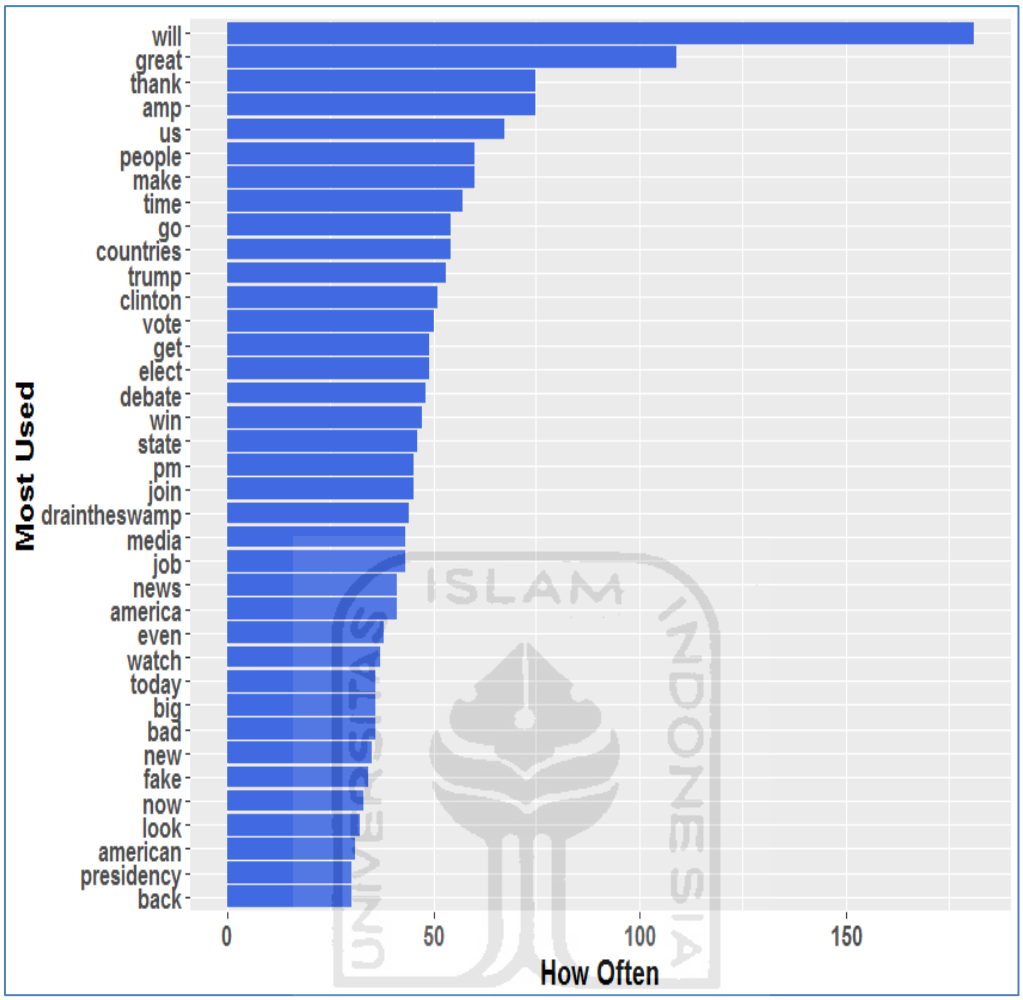
- trump_plot<- ggplot(df, aes(x = reorder(term, freq), y = freq)) + geom_bar(stat = "identity", fill = "royalblue") + xlab("Most Used") + ylab("How Often") + coord_flip() + theme(text=element_text(size=15,face="bold"))
- trump_plot

Didapatkan *output* sebagai berikut:



Gambar 5.2.1.1. Grafik *Term Frequency* Akun @BarakObama

Term Frequency menampilkan kata-kata yang seringkali keluar dari data teks yang dianalisis. Kata-kata yang sering keluar kemudian divisualisasikan dalam bentuk diagram batang yang menunjukkan frekuensi kata tersebut keluar. Semakin panjang gambar batangnya, berarti semakin banyak kata tersebut keluar. Dari gambar 5.2.1.1. di atas, dapat diketahui bahwa kata yang paling banyak keluar adalah "senate". Kata "senate" merupakan kata yang paling banyak di tulis oleh @BarackObama, disusul oleh "obama" dan "president". Kemudian kata-kata yang paling banyak keluar selanjutnya adalah "doyourjob", "leader", "actonclimate", "judge", "garland", "supreme", "court", "climate", "now", "change", "american" dan "hear".



Gambar 5.2.1.2. Grafik *Term Frequency* Akun @RealDonaldTrump

Dari gambar 5.2.1.2 di atas, dapat diketahui bahwa kata yang paling banyak keluar adalah “will”. Kata “will” merupakan kata yang paling banyak di tulis oleh @realDonaldTrump, disusul oleh “great”. Kemudian kata-kata yang paling banyak keluar selanjutnya adalah “thank”, “amp”, “us”, “people”, “make”, “time”, “go”, “countries”, “trump”, “clinton”, “vote”, get”, “elect”, “debate”, “win”, ”state”, “pm”, “join”, “draintheswamp”, “media”, “job”, “news”, “america”, “even”, “watch”, “today”, “big”, “bad”, “new”, “fake”, “now”, “look”, “american”, “presidency”, dan “back”.

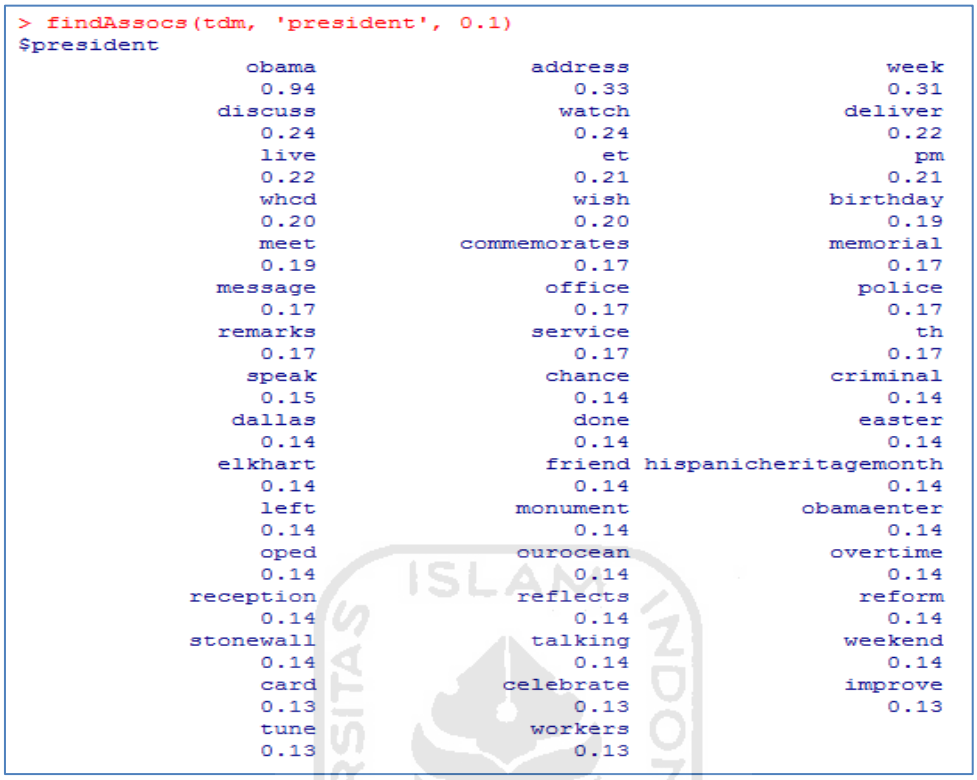
```
> findAssocs(tdm, 'senate', 0.1)
$senate
          leader       doyourjob           court         supreme
            0.74            0.68            0.42            0.42
        garland           judge            fill         vacancy
            0.40            0.39            0.36            0.34
            give            hear        obstruct            fair
            0.30            0.30            0.30            0.26
       editorial  qualifications         refusal          scotus
            0.23            0.23            0.23            0.23
           block          boards        consider             job
            0.21            0.21            0.21            0.21
       nominated            call       political             put
            0.21            0.19            0.19            0.19
  unprecedented            tell     republicans             six
            0.19            0.18            0.17            0.17
            high    unacceptable            vote         nominee
            0.16            0.16            0.15            0.14
            time       dutiessay            even            hold
            0.14            0.13            0.13            0.13
         merrick     obstruction obstructionists            oped
            0.13            0.13            0.13            0.13
        partisan       precedent        pressure         process
            0.13            0.13            0.13            0.13
       professor         recess          across        continue
            0.13            0.13            0.11            0.11
           biden            stop            wait
            0.10            0.10            0.10
```

Gambar 5.2.1.3. Peluang Kata yang Berasosiasi dengan Kata "Senate" dari Akun @BarackObama

Perintah "*findAssocs(tdm, 'senate', 0.1)*" seperti pada gambar 5.2.1.3. berfungsi untuk melihat nilai korelasi yang berhubungan dengan kata "*senate*" sebesar 0,1. Peneliti menggunakan angka 0,1 sebagai angka keterhubungan terkecil sehingga dengan itu dapat dilihat kata-kata yang berhubungan dari terkecil hingga terbesar. Hasil *output* yang dihasilkan menunjukkan bahwa kata-kata yang berhubungan dengan kata "*senate*" adalah "leader", kemudian disusul oleh "doyourjob", "court", "supreme", "garland", "judge", "fill", "facancy", "give", "hear", "abstruct", "fair", "editorial", "qualifications", "refusal", "scotus", "block", "boards", "consider", "job", "nominated", "call", "political", "put", "unprecedented", "tell", "republicans", "six", "high", "unacceptable", "vote", "nominee", "time", "dutiessay", "even", "hold", "merrick", "obstruction", "obstructionists", "oped", "partisan", "precedent", "pressure", "process", "professor", "recess", "across", "continue", "biden", "stop", "wait".

Dengan besar asosiasi tertinggi "senate" dengan "leader" sebesar 0.74; kemudian disusul oleh "senate" dengan "doyourjob" sebesar 0.68; "senate"

dengan “court” sebesar 0.42; “senate” dengan “supreme” sebesar 0.42; “senate” dengan “garland” sebesar 0.40; “senate” dengan “judge” sebesar 0.39; “senate” dengan “fill” 0.36; “senate” dengan “facancy” sebesar 0.34; “senate” dengan “give” sebesar 0.30; “senate” dengan “hear” sebesar 0.30; “senate” dengan “abstruct” sebesar 0.30; “senate” dengan “fair” sebesar 0.26; “senate” dengan “editorial” sebesar 0.23; “senate” dengan “qualifications” sebesar 0.23; “senate” dengan “refusal” sebesar 0.23; “senate” dengan “scotus” sebesar 0.23; “senate” dengan “block” sebesar 0.21; “senate” dengan “boards” sebesar 0.21; “senate” dengan “consider” sebesar 0.21; “senate” dengan “job” sebesar 0.21; “senate” dengan “nominated” sebesar 0.21; “senate” dengan “call” sebesar 0.19; “senate” dengan “political” sebesar 0.19; “senate” dengan “put” sebesar 0.19; “senate” dengan “unprecedented” sebesar 0.19; “senate” dengan “tell” sebesar 0.18; “senate” dengan “republicans” sebesar 0.17; “senate” dengan “six” sebesar 0.17; “senate” dengan “high” sebesar 0.16; “senate” dengan “unacceptable” sebesar 0.16; “senate” dengan “vote” sebesar 0.15; “senate” dengan “nominee” sebesar 0.14; “senate” dengan “time” sebesar 0.14; “senate” dengan “dutiessay” sebesar 0.13; “senate” dengan “even” sebesar 0.13; “senate” dengan “hold” sebesar 0.13; “senate” dengan “merrick” sebesar 0.13; “senate” dengan “obstruction” sebesar 0.13; “senate” dengan “obstructionists” sebesar 0.13; “senate” dengan “oped” sebesar 0.13; “senate” dengan “partisan” sebesar 0.13; “senate” dengan “precedent” sebesar 0.13; “senate” dengan “pressure” sebesar 0.13; “senate” dengan “process” sebesar 0.13; “senate” dengan “professor” sebesar 0.13; “senate” dengan “recess” sebesar 0.13; “senate” dengan “across” sebesar 0.11; “senate” dengan “continue” sebesar 0.11; “senate” dengan “biden” sebesar sebesar 0.10; “senate” dengan “stop” sebesar 0.10; “senate” dengan “wait” sebesar 0.10.

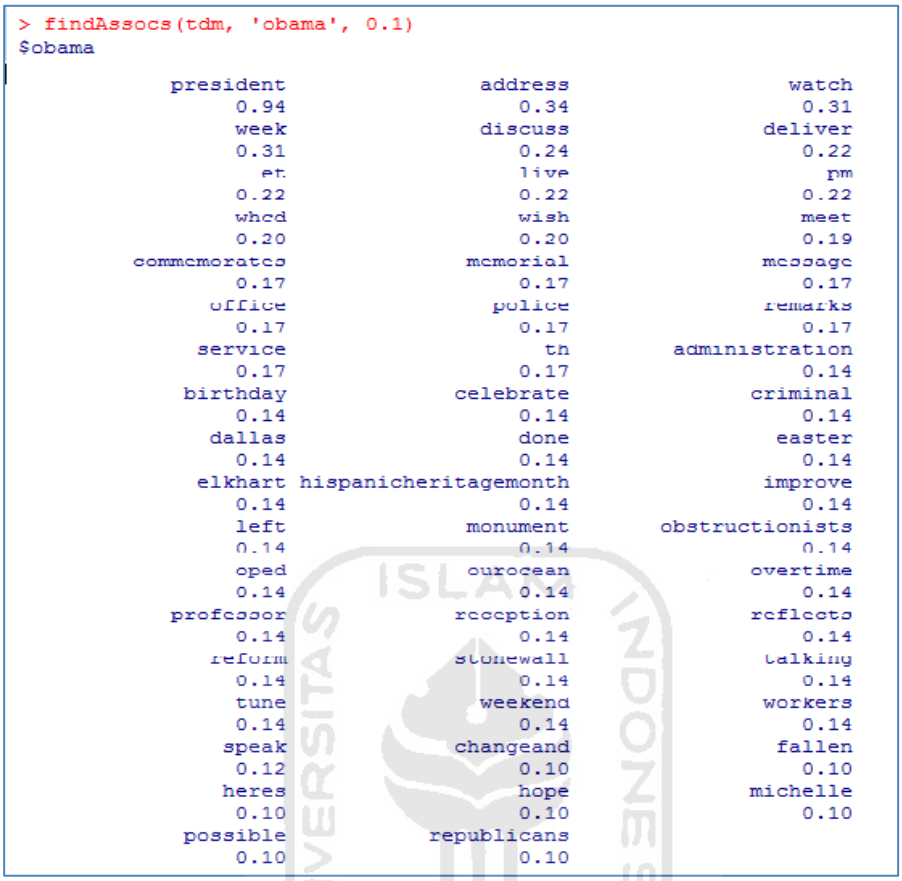


```
> findAssocs(tdm, 'president', 0.1)
$president
    obama   address      week 
     0.94      0.33      0.31 
  discuss     watch   deliver 
     0.24      0.24      0.22 
     live        et        pm 
     0.22      0.21      0.21 
     whcd      wish  birthday 
     0.20      0.20      0.19 
     meet commemorates  memorial 
     0.19      0.17      0.17 
  message    office    police 
     0.17      0.17      0.17 
  remarks   service        th 
     0.17      0.17      0.17 
    speak    chance  criminal 
     0.15      0.14      0.14 
   dallas      done    easter 
     0.14      0.14      0.14 
  elkhart    friend hispanicheritagemonth 
     0.14      0.14      0.14 
     left  monument  obamaenter 
     0.14      0.14      0.14 
     oped  ourocean  overtime 
     0.14      0.14      0.14 
reception   reflects    reform 
     0.14      0.14      0.14 
stonewall   talking   weekend 
     0.14      0.14      0.14 
     card  celebrate   improve 
     0.13      0.13      0.13 
     tune   workers 
     0.13      0.13 
```

Gambar 5.2.1.4. Peluang Kata yang Berasosiasi dengan Kata "President" dari Akun @BarackObama

Perintah "*findAssocs(tdm, 'president', 0.1)*" seperti pada gambar 5.2.1.4. berfungsi untuk melihat nilai korelasi yang berhubungan dengan kata "*president*" sebesar 0,1. Peneliti menggunakan angka 0,1 sebagai angka keterhubungan terkecil sehingga dengan itu dapat dilihat kata-kata yang berhubungan dari terkecil hingga terbesar. Hasil *output* yang dihasilkan menunjukkan bahwa kata-kata yang berhubungan dengan kata "*president*" adalah "obama", kemudian disusul oleh "address", "week", "discuss", "watch", "deliver", "live", "et", "pm", "whcd", "wish", "birthday", "meet", "commemorates", "memorial", "message", "office", "police", "remarks", "service", "th", "speak", "chance", "criminal", "dallas", "done", "easter", "elkhart", "friend", "hispaniceritagemonth", "left", "monument", "obamaenter", "oped", "ourocean", "overtime", "reception", "reflects", "reform", "stonewall", "talking", "weekend", "card", "celebrate", "improve", "tune", "workers".

Dengan besar asosiasi tertinggi “president” dengan “obama” sebesar 0.94. Kemudian disusul oleh “president” dengan ”address” sebesar 0.33; “president” dengan “week” sebesar 0.31; “president” dengan “discuss” sebesar 0.24; “president” dengan “watch” sebesar 0.24; “president” dengan “deliver” sebesar 0.22; “president” dengan “live” sebesar 0.22; “president” dengan “et” sebesar 0.21; “president” dengan “pm” sebesar 0.21; “president” dengan “whcd” sebesar 0.20; “president” dengan “wish” sebesar 0.20; “president” dengan “birthday” sebesar 0.19; “president” dengan “meet” sebesar 0.17; “president” dengan “commemorates” sebesar 0.17; “president” dengan “memorial” sebesar 0.17; “president” dengan “message” sebesar 0.17; “president” dengan “office” sebesar 0.17; “president” dengan “police” sebesar 0.17; “president” dengan “remarks” sebesar 0.17; “president” dengan “service” sebesar 0.17; “president” dengan “th” sebesar 0.17; “president” dengan “speak” sebesar 0.15; “president” dengan “chance” sebesar 0.14; “president” dengan “criminal” sebesar 0.14; “president” dengan “dallas” sebesar 0.14; “president” dengan “done” sebesar 0.14; “president” dengan “easter” sebesar 0.14; “president” dengan “elkhart” sebesar 0.14; “president” dengan “friend” sebesar 0.14; “president” dengan “hispaniceritagemonth” sebesar 0.14; “president” dengan “left” sebesar 0.14; “president” dengan “monument” sebesar 0.14; “president” dengan “obamaenter” sebesar 0.14; “president” dengan “oped” sebesar 0.14; “president” dengan “ourocean” sebesar 0.14; “president” dengan “overtime” sebesar 0.14; “president” dengan “reception” sebesar 0.14; “president” dengan “reflects” sebesar 0.14; “president” dengan “reform” sebesar 0.14; “president” dengan “stonewall” sebesar 0.14; “president” dengan “talking” sebesar 0.14; “president” dengan “weekend” sebesar 0.14; “president” dengan “card” sebesar 0.13; “president” dengan “celebrate” sebesar 0.13; “president” dengan “improve” sebesar 0.13; “president” dengan “tune” sebesar 0.13; “president” dengan “workers” sebesar 0.13.



```
> findAssocs(tdm, 'obama', 0.1)
$obama
            president               address                 watch
                 0.94                  0.34                  0.31
                 week               discuss               deliver
                 0.31                  0.24                  0.22
                   et                  live                    pm
                 0.22                  0.22                  0.22
                 whcd                  wish                  meet
                 0.20                  0.20                  0.19
        commemorates              memorial               message
                 0.17                  0.17                  0.17
               office                police               remarks
                 0.17                  0.17                  0.17
              service                    th        administration
                 0.17                  0.17                  0.14
            birthday             celebrate              criminal
                 0.14                  0.14                  0.14
              dallas                  done                easter
                 0.14                  0.14                  0.14
             elkhart hispanicheritagemonth               improve
                 0.14                  0.14                  0.14
               left              monument      obstructionists
                 0.14                  0.14                  0.14
               oped              ourocean              overtime
                 0.14                  0.14                  0.14
          professor             reception              reflects
                 0.14                  0.14                  0.14
             reform             stonewall               talking
                 0.14                  0.14                  0.14
               tune               weekend               workers
                 0.14                  0.14                  0.14
              speak             changeand                fallen
                 0.12                  0.10                  0.10
              heres                  hope               michelle
                 0.10                  0.10                  0.10
           possible           republicans
                 0.10                  0.10
```

Gambar 5.2.1.5. Peluang Kata yang Berasosiasi dengan Kata "Obama" dari Akun @BarackObama

Perintah "*findAssocs(tdm, 'president', 0.1)*" seperti pada gambar 5.2.1.5. berfungsi untuk melihat nilai korelasi yang berhubungan dengan kata "obama" sebesar 0,1. *Output* yang dihasilkan menunjukkan bahwa kata-kata yang berhubungan dengan kata "*obama*" adalah "president", kemudian disusul oleh "address", "watch", "week", "discuss", "deliver", "et", "live", "pm", "whcd", "wish", "meet", "commemorate", "memorial", "message", "office", "police", "remarks", "service", "th", "administration", "birthday", "celebrate", "criminal", "dallas", "done", "easter", "elkhart", "hispaniceritagemonth", "improve", "left", "monument", "obstructionists", "oped", "ourocean", "overtime", "professor", "reception", "refiects", "reform", "stonewall", "talking", "tune", "weekend", "workers", "speak", "changeand", "fallen", "heres", "hope", "michele", "possible", "republicans".

Dengan besar asosiasi tertinggi “obama” dengan “president” sebesar 0.94; Kemudian disusul oleh “president” dengan“address” sebesar 0.34; “president” dengan “watch” sebesar 0.31; “president” dengan “week” sebesar 0.31; “president” dengan “discuss” sebesar 0.24; “president” dengan “deliver” sebesar 0.22; “president” dengan “et” sebesar 0.22; “president” dengan “live” sebesar 0.22; “president” dengan “pm” sebesar 0.22; “president” dengan “whcd” sebesar 0.20; “president” dengan “wish” sebesar 0.20; “president” dengan “meet” sebesar 0.19; “president” dengan “commemorate” sebesar 0.17; “president” dengan “memorial” sebesar 0.17; “president” dengan “message” sebesar 0.17; “president” dengan “office” sebesar 0.17; “president” dengan “police” sebesar 0.17; “president” dengan “remarks” sebesar 0.17; “president” dengan “service” sebesar 0.17; “president” dengan “th” sebesar 0.17; “president” dengan “administration” sebesar 0.14; “president” dengan “birthday” sebesar 0.14; “president” dengan “celebrate” sebesar 0.14; “president” dengan “criminal” sebesar 0.14; “president” dengan “dallas” sebesar 0.14; “president” dengan “done” sebesar 0.14; “president” dengan “easter” sebesar 0.14; “president” dengan “elkhart” sebesar 0.14; “president” dengan “hispaniceritagemonth” sebesar 0.14; “president” dengan “improve” sebesar 0.14; “president” dengan “left” sebesar 0.14; “president” dengan “monument” sebesar 0.14; “president” dengan “obstructionists” sebesar 0.14; “president” dengan “oped” sebesar 0.14; “president” dengan “ourocean” sebesar 0.14; “president” dengan “overtime” sebesar 0.14; “president” dengan “professor” sebesar 0.14; “president” dengan “reception” sebesar 0.14; “president” dengan “refiects” sebesar 0.14; “president” dengan “reform” sebesar 0.14; “president” dengan “stonewall” sebesar 0.14; “president” dengan “talking” sebesar 0.14; “president” dengan “tune” sebesar 0.14; “president” dengan “weekend”; “president” dengan “workers” sebesar 0.14; “president” dengan “speak” sebesar 0.12; “president” dengan “changeand” sebesar 0.10; “president” dengan “fallen” sebesar 0.10; “president” dengan “heres” sebesar 0.10; “president” dengan”hope” sebesar 0.10; “president” dengan “michele” sebesar 0.10; “president” dengan “possible” sebesar 0.10; “president” dengan “republicans” sebesar 0.10.

```
> findAssocs(tdm, 'will', 0.1)
$will
         bring           make      slaughter           soon          speak           back         wealth
          0.24           0.19           0.18           0.18           0.18           0.17           0.17
     interview        america          great       together           wife          doubt          dream
          0.15           0.14           0.14           0.14           0.14           0.13           0.13
         eric godblesstheusa         killed           pigs          swear        weekend administration
          0.13           0.13           0.13           0.13           0.13           0.13           0.12
      anderson         anyway        ballots         cooper          crush           deep            don
          0.12           0.12           0.12           0.12           0.12           0.12           0.12
        enjoy      formation        harder         manage        melania         oclock       peacetoo
          0.12           0.12           0.12           0.12           0.12           0.12           0.12
       triple     untrusting         weight             wh       children     forgotten
          0.12           0.12           0.12           0.12           0.11           0.10
```

Gambar 5.2.1.6. Peluang Kata yang Berasosiasi dengan Kata "Will" dari Akun @realDonaldTrump

Perintah "*findAssocs(tdm, 'will', 0.1)*" seperti pada gambar 5.2.1.6. berfungsi untuk melihat nilai korelasi yang berhubungan dengan kata "will" sebesar 0,1. *Output* yang dihasilkan menunjukkan bahwa kata-kata yang berhubungan dengan kata "will" adalah "bring", kemudian disusul oleh "make", "slaughter", "soon", "speak", "back", "wealth", "interview", "america", "great", "together", "wife", "doubt", "dream", "eric", "godblesstheusa", "killed", "pigs", "swear", "weekend", "administration", "anderson", "anyway", "ballots", "cooper", "crush", "deep", "don", "enjoy", "formation", "harder", "manage", "melania", "oclock", "peacetoo", "triple", "untrusting", "weight", "wh", "children", "forgotten".

Dengan besar asosiasi tertinggi "will" dengan "bring" sebesar 0.24. Kemudian disusul oleh "will" dengan "make" sebesar 0.19; "will" dengan "slaughter" sebesar 0.18; "will" dengan "soon" sebesar 0.18; "will" dengan "speak" sebesar 0.18; "will" dengan "back" sebesar 0.17; "will" dengan "wealth" sebesar 0.17; "will" dengan "interview" sebesar 0.15; "will" dengan "america" sebesar 0.14; "will" dengan "great" sebesar 0.14; "will" dengan "togethe" sebesar 0.14; "will" dengan "wife" sebesar 0.14; "will" dengan "doubt" sebesar 0.13; "will" dengan "dream" sebesar 0.13; "will" dengan "eric" sebesar 0.13; "will" dengan "godblesstheusa" sebesar 0.13; "will" dengan "killed" sebesar 0.13; "will" dengan "pigs" sebesar 0.13; "will" dengan "swear" sebesar 0.13; "will" dengan "weekend" sebesar 0.13; "will" dengan "administration" sebesar 0.12; "will" dengan "anderson" sebesar 0.12; "will" dengan "anyway" sebesar 0.12; "will"

dengan “ballots” sebesar 0.12; “will” dengan “cooper” sebesar 0.12; “will” dengan “crush” sebesar 0.12; “will” dengan “deep” sebesar 0.12; “will” dengan “don” sebesar 0.12; “will” dengan “enjoy” sebesar 0.12; “will” dengan “formation” sebesar 0.12; “will” dengan “herder” sebesar 0.12; “will” dengan “manage” sebesar 0.12; “will” dengan “melania” sebesar 0.12; “will” dengan “oclock” sebesar 0.12; “will” dengan “peacetoo” sebesar 0.12; “will” dengan “triple” sebesar 0.12; “will” dengan “untrusting” sebesar 0.12; “will” dengan “weight” sebesar 0.12; “will” dengan “wh” sebesar 0.12; “will” dengan “children” sebesar 0.11; “will” dengan “forgotten” sebesar 0.10.

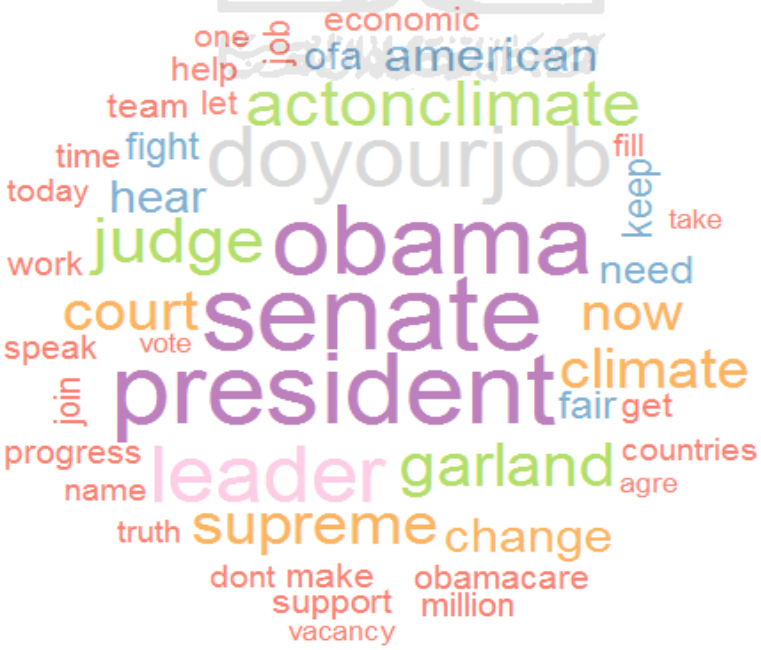
```
> findAssocs(tdm, 'great', 0.1)
$great
     america         make     together   ambassador       ivanka       kasich          peo      britain
        0.40         0.30         0.22         0.18         0.18         0.18         0.18         0.17
   dignified headquarters    insurance  nigelfarage     ovations        owens       packed      provide
        0.17         0.17         0.17         0.17         0.17         0.17         0.17         0.17
      people       attend          ceo    fantastic         will         john         meet       person
        0.15         0.14         0.14         0.14         0.14         0.13         0.13         0.13
       state      amazing     daughter          guy         paid       become        brave        honor
        0.13         0.12         0.12         0.12         0.12         0.11         0.11         0.11
        joke    listening    represent        thank      weekend    yesterday
        0.11         0.11         0.11         0.11         0.11         0.10
```

Gambar 5.2.1.7. Peluang Kata yang Berasosiasi dengan Kata “Great” dari Akun @realDonaldTrump

Perintah “*findAssocs(tdm, ‘great’, 0.1)*” seperti pada gambar 5.2.1.7. berfungsi untuk melihat nilai korelasi yang berhubungan dengan kata “great” sebesar 0,1. *Output* yang dihasilkan menunjukkan bahwa kata-kata yang berhubungan dengan kata “great” adalah “america”, kemudian disusul oleh “make”, “together”, “ambassador”, “ivanka”, “kasich”, “peo”, “britain”, “dignified”, “headquarters”, “insurance”, “nigelfarage”, “ovations”, “owens”, “packed”, “provide”, “people”, “attend”, “ceo”, “fantastic”, “will”, “john”, “meet”, “person”, “state”, “amazing”, “daughter”, “guy”, “paid”, “become”, “brave”, “honor”, “joke”, “listening”, “represent”, “thank”, “weekend”, “yesterday”.

Dengan besar asosiasi tertinggi “great” dengan “america” sebesar 0.40. Kemudian disusul oleh “great” dengan “make” sebesar 0.30; “great” dengan “together” sebesar 0.22; “great” dengan “ambassador” sebesar 0.18; “great”

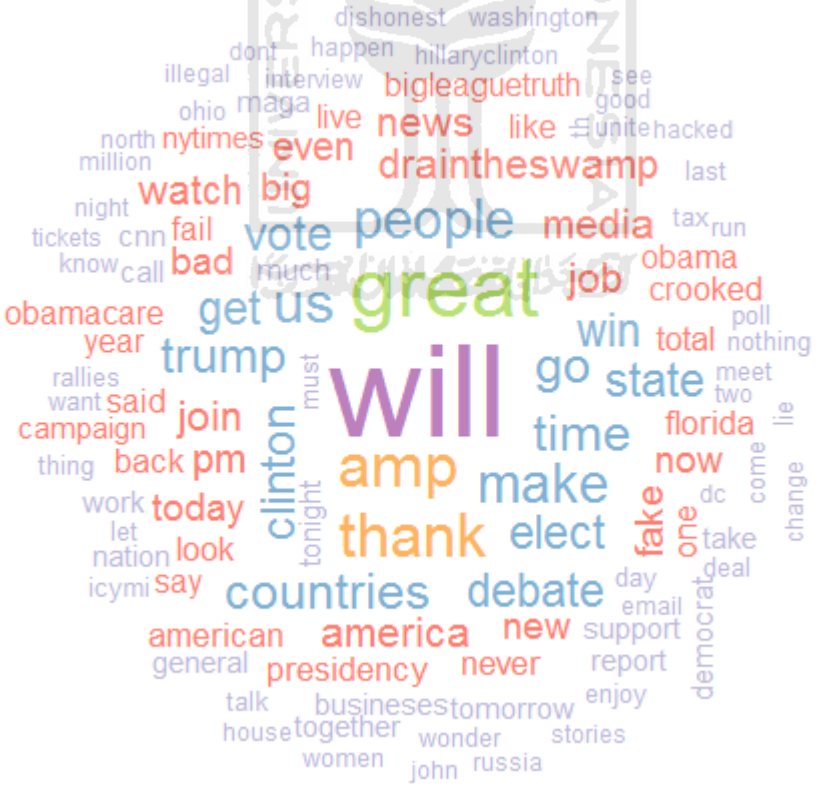
dengan “ivanka” sebesar 0.18; “great” dengan “kasich” sebesar 0.18; “great” dengan “peo” sebesar 0.18; “great” dengan “britain” sebesar 0.17; “great” dengan “dignified” sebesar 0.17; “great” dengan “headquarters” sebesar 0.17; “great” dengan “insurance” sebesar 0.17; “great” dengan “nigelfarage” sebesar 0.17; “great” dengan “ovations” sebesar 0.17; “great” dengan “owens” sebesar 0.17; “great” dengan “packed” sebesar 0.17; “great” dengan “provide” sebesar 0.17; “great” dengan “people” sebesar 0.15; “great” dengan “attend” sebesar 0.14; “great” dengan “ceo” sebesar 0.14; “great” dengan “fantastic” sebesar 0.14; “great” dengan “will” sebesar 0.14; “great” dengan “john” sebesar 0.13; “great” dengan “meet” sebesar 0.13; “great” dengan “person” sebesar 0.13; “great” dengan “state” sebesar 0.13; “great” dengan “amazing” sebesar 0.12; “great” dengan “daughter” sebesar 0.12; “great” dengan “guy” sebesar 0.12; “great” dengan “paid” sebesar 0.12; “great” dengan “become” sebesar 0.11; “great” dengan “brave” sebesar 0.11; “great” dengan “honor” sebesar 0.11; “joke” sebesar 0.11; “listening” sebesar 0.11; “great” dengan “represent” sebesar 0.11; “great” dengan “thank” sebesar 0.11; “great” dengan “weekend” sebesar 0.11; “great” dengan “yesterday” sebesar 0.10.



Gambar 5.2.1.8. Word Cloud Akun @BarackObama

Word cloud atau awan kata merupakan kumpulan kata-kata yang paling banyak keluar dari data teks yang dianalisis. Kata-kata tersebut berkumpul seperti sebuah awan yang berisi kata-kata sehingga disebut awan kata. Semakin besar huruf dari kata yang terdapat di awan kata, menunjukkan semakin sering kata tersebut keluar. Tampilan awan kata ini lebih menarik karena lebih mudah untuk dilihat secara keseluruhan, tetapi tidak sedetail diagram frekuensi dalam menunjukkan frekuensi kata-kata keluar dalam suatu teks yang dianalisis.

Dari awan kata pada gambar 5.2.1.8., dapat diketahui bahwa kata atau topik yang paling sering dikicaukan oleh Barack Obama adalah “senate”, selanjutnya “president”, “obama”, dan sebagainya. Sedangkan pada awan kata gambar 5.2.1.9., dapat diketahui bahwa kata atau topik yang paling sering dikicaukan oleh Donald Trump adalah “will”, selanjutnya “great” dan sebagainya.



Gambar 5.2.1.9. Word Cloud Akun @realDonaldTrump

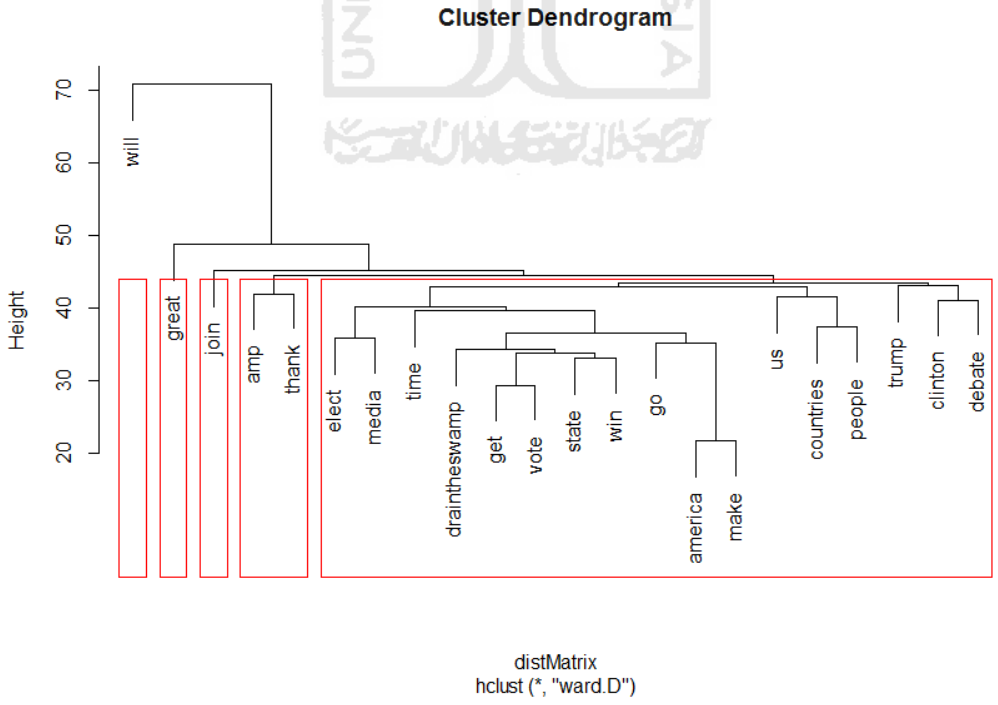
ISLAM
UNIVERSITAS
INDONESIA

Berdasarkan gambar 5.2.2.1. didapatkan pembagian kelompok kata untuk akun @BarackObama sebagai berikut:

a. Kelompok 1: *actonclimate, change, climate*
b. Kelompok 2: *american, fight, get, keep, make, need, now, ofa, support*
c. Kelompok 3: *court, doyourjob, leader, senate, supreme*
d. Kelompok 4: *fair, garland, hear, judge*
e. Kelompok 5: *obama, president*

Berdasarkan gambar 5.2.2.2. didapatkan pembagian kelompok kata untuk akun @BarackObama sebagai berikut:

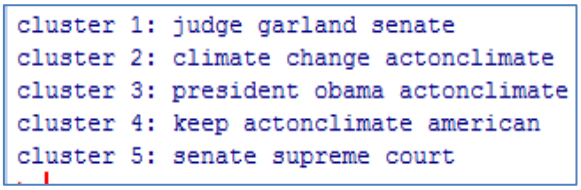
a. Kelompok 1: *america, clinton, countries, debate, draintheswamp, elect, get, go, make, media, people, state, time, trump, us, vote, win*
b. Kelompok 2: *amp, thank*
c. Kelompok 3: *great*
d. Kelompok 4: *join*
e. Kelompok 5: *will*



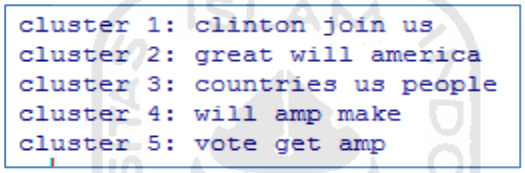
Gambar 5.2.2.2. Dendogram akun @realDonaldTrump

5.2.3. Perbedaan Topik Pembicaraan

Clustering tweet digunakan untuk melihat kelompok *tweet* berdasarkan topik-topik yang berbeda tetapi memilik karakteristik yang sama. Berikut ini merupakan hasil pengelompkan *tweet* menggunakan metode *K-Means* untuk akun @BarackObama dan @realDonaldTrump.



Gambar 5.2.3.1. *Cluster K-Means* akun @BarackObama

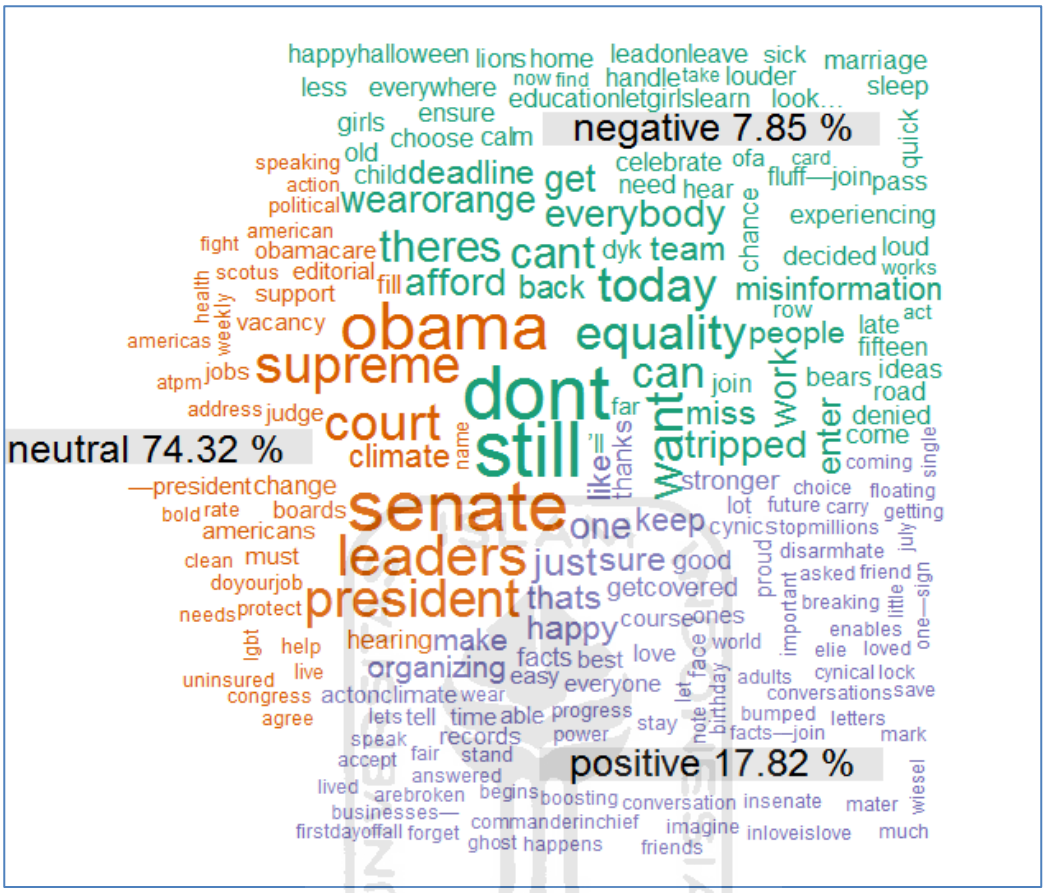


Gambar 5.2.3.2. *Cluster K-Means* akun @realDonaldtrump

Gambar 5.2.3.1. dan 5.2.3.2. diatas adalah hasil cluster tweets untuk masing-masing akun (@BarackObama dan @realDonaldTrump), dimana apabila diperhatikan tidak terdapat kesamaan topik pembicaraan. Topik pembicaraan dapat dilihat dari kata pertama pada setiap kelompoknya. Topik pembicaraan yang sering dicuitkan oleh @BarackObama adalah *judge*, *climate*, *president, keep*, dan *senate*. Sedangkan pada @realDonaldTrump adalah *clinton*, *great*, *coutries*, *will* dan *vote*.

5.2.4. Analisis sentimen

Pada tahap pengujian sistem akan mencari nilai klasifikasi terhadap data tweet secara otomatis. Nilai yang diperoleh berupa sentimen *positive*, sentimen *negative* dan sentimen *neutral* dalam bentuk *wordcloud*. Pengujian ini menggunakan form klasifikasi dengan cara jumlah data yang akan di analisa, yaitu 2000 tweets untuk masing-masing akun.



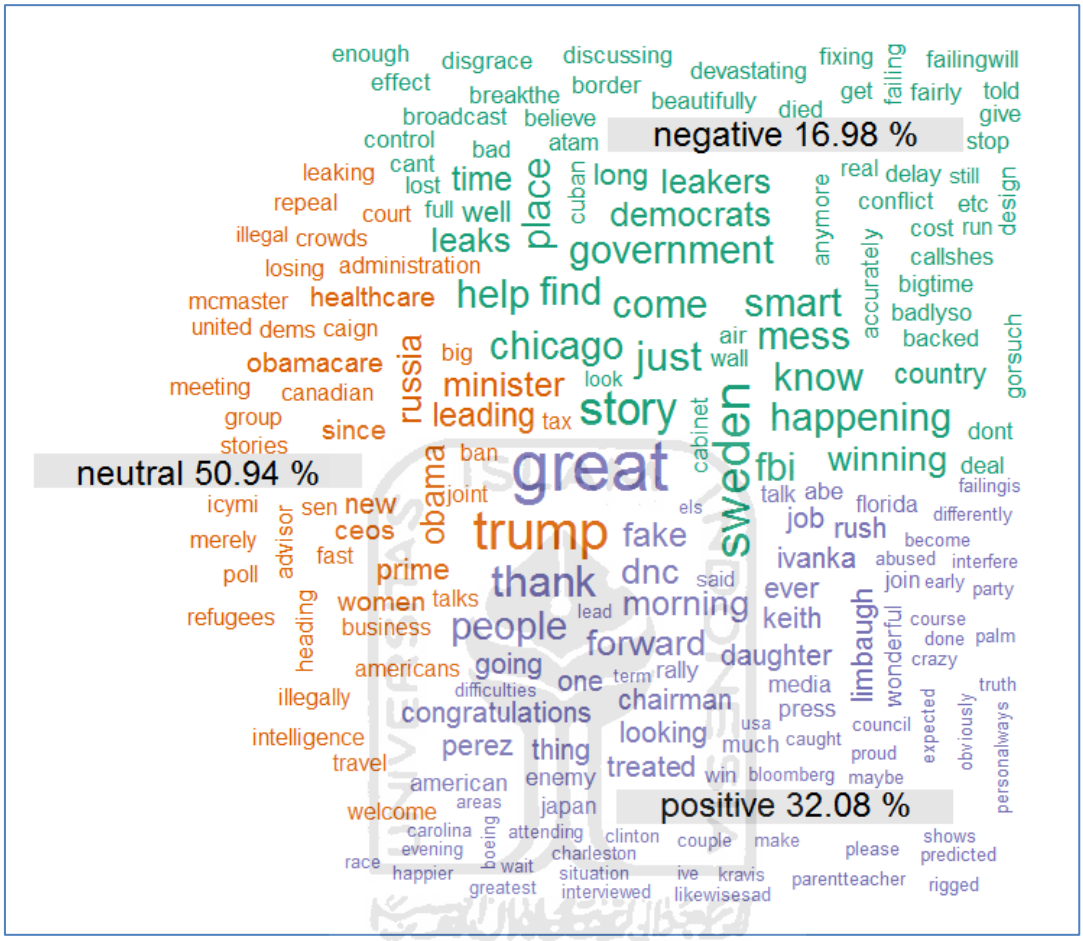
Gambar 5.2.4.1. *Wordcloud Sentiment Analysis* @BarackObama

Pada gambar 5.2.4.1. dapat dilihat cuitan-cuitan dari akun @BarackObama lebih kepada cuitan yang netral, lalu positif dan paling kecil adalah negatif. Dari hasil uji analisis sentimen pada akun *twitter Barack Obama* didapat *tweet* dengan tiga kategori yaitu kategori positif sebesar 17.82%, negatif 7.85% dan netral 74.32%.

Sedangkan pada gambar 5.2.4.2. dapat dilihat cuitan-cuitan dari akun @realDonaldTrump lebih kepada cuitan yang netral, lalu positif dan paling kecil adalah negatif. Dari hasil uji analisis sentimen pada akun *twitter Donald Trump* didapat *tweet* dengan tiga kategori yaitu kategori positif sebesar 32.08%, negatif 16.98% dan netral 50.94%.

Jika dibandingkan dari dua akun tersebut, maka didapatkan kesimpulan dari besaran tiap kategori, *Barack Obama* lebih banyak berbicara netral

dibandingkan dengan *Donald Trump* karena angka cuitan netral *Obama* (74.32%) lebih besar dibandingkan dengan angka cuitan netral *Trump* (50.94%).



Gambar 5.2.4.2. *Wordcloud Sentiment Analysis* @realDonaldTrump

BAB VI

KESIMPULAN DAN SARAN

6.1. Kesimpulan

Berdasarkan analisis *text mining* yang dilakukanuntuk penanganan data besar hasil pencarian topik-topik terkait pada akun *twitter* @BarackObama dan @realDonaldTrump, maka peneliti dapat menarik beberapa kesimpulan antara lain:

a. Topik utama yang melekat pada Barack Obama adalah *senate*, sedangkan topik utama yang melekat pada Donald Trump adalah *will*.

b. Topik maupun kata lainnya yang melekat/saling berhubungan pada Barack Obama adalah *obama, president,* doyourjob, *leader, actonclimate, judge, garland, supreme, court, climate, now, change, american,* dan *hear.*
Topik maupun kata lainnya yang melekat/saling berhubungan pada Donald Trump adalah *great, thank, amp, us, people, make, time, go, countries, trump, clinton, vote, get, elect, debate, win, state, pm, join, draintheswamp, media, job, news, america, even, watch, today, big, bad, new, fake, now, look, american, presidency,* dan *back.*

c. Pada akun *twitter* Barack Obama dengan menggunakan ukuran asosiasi data dengan nilai korelasi tidak kurang dari 0.10 untuk kata *senate adalah leader, doyourjob, court, supreme, garland, judge, fill, facancy, give, hear, abstruct, fair, editorial, qualifications, refusal, scotus, block, boards, consider, job, nominated, call, political, put, unprecedented, tell, republicans, six, high, unacceptable, vote, nominee, time, dutiessay, even, hold, merrick, obstruction, obstructionists, oped, partisan, precedent, pressure, process, professor, recess, across, continue, biden, stop,* dan *wait.* Untuk kata *president* adalah *obama, address, week, discuss, watch, deliver, live, et, pm, whcd, wish, birthday, meet, commemorates, memorial, message, office, police, remarks, service, th, speak, chance, criminal,*

dallas, done, easter, elkhart, friend, hispaniceritagemonth, left, monument, obamaenter, oped, ourocean, overtime, reception, reflects, reform, stonewall, talking, weekend, card, celebrate, improve, tune, dan *workers*. Untuk kata *obama* adalah *president, address, live, deliver, watch, week, discuss, administration, th,* dan *wish.* Dan untuk kata *president adalah obama, address, watch, week, discuss, deliver, et, live, pm, whcd, wish, meet, commemorate, memorial, message, office, police, remarks, service, th, administration, birthday, celebrate, criminal, dallas, done, easter, elkhart, hispaniceritagemonth, improve, left, monument, obstructionists, oped, ourocean, overtime, professor, reception, refiects, reform, stonewall, talking, tune, weekend, workers, speak, changeand, fallen, heres, hope, michele, possible,* dan *republicans.*

d. Pada akun *twitter* Donald Trump dengan menggunakan ukuran asosiasi data dengan nilai korelasi tidak kurang dari 0.10 untuk kata *will* adalah *bring, make, slaughter, soon, speak, back, wealth, interview, america, great, together, wife, doubt, dream, eric, godblesstheusa, killed, pigs, swear, weekend, administration, anderson, anyway, ballots, cooper, crush, deep, don, enjoy, formation, harder, manage, melania, oclock, peacetoo, triple, untrusting, weight, wh, children,* dan *forgotten.* Dan untuk kata *great* didapatkan kata *america, make, together, ambassador, ivanka, kasich, peo, britain, dignified, headquarters, insurance, nigelfarage, ovations, owens, packed, provide, people, attend, ceo, fantastic, will, john, meet, person, state, amazing, daughter, guy, paid, become, brave, honor, joke, listening, represent, thank, weekend,* dan *yesterday.*

e. Dengan menggunakan *cluster k means* dibagi 5 kelompok tweet pada masing-masing akun. Untuk akun twitter Barack Obama didapatkan kelompok sebagai berikut:

   Kelompok 1: *actonclimate, change, climate*

   Kelompok 2: *american, fight, get, keep, make, need, now, ofa, support*

   Kelompok 3: *court, doyourjob, leader, senate, supreme*

Kelompok 4: *fair, garland, hear, judge*

Kelompok 5: *obama, president*

Dan untuk akun twitter Donald Trump didapatkan kelompok sebagai berikut:

Kelompok 1: *america, clinton, countries, debate, draintheswamp, elect, get, go, make, media, people, state, time, trump, us, vote, win*

Kelompok 2: *amp, thank*

Kelompok 3: *great*

Kelompok 4: *join*

Kelompok 5: *will*

f. Dari hasil cluster tweets untuk masing-masing akun (@BarackObama dan @realDonaldTrump), tidak terdapat kesamaan topik pembicaraan. Topik pembicaraan yang sering dicuitkan oleh @BarackObama adalah *judge, climate, president, keep,* dan *senate*. Sedangkan pada @realDonaldTrump adalah *clinton, great, coutries, will* dan *vote*.

g. Dari hasil uji analisis sentimen pada akun *twitter Barack Obama* didapat *tweet* dengan tiga kategori yaitu kategori positif sebesar 17.82%, negatif 7.85% dan netral 74.32%. Sedangkan pada akun *twitter Donald Trump* didapat *tweet* dengan tiga kategori yaitu kategori positif sebesar 32.08%, negatif 16.98% dan netral 50.94%.

   Berdasarkan analisis sentimen *Barack Obama* lebih banyak berbicara netral dibandingkan dengan *Donald Trump* karena angka cuitan netral *Obama* (74.32%) lebih besar dibandingkan dengan angka cuitan netral *Trump* (50.94%).

6.2. Saran

Berdasarkan pengujian yang telah dilakukan, terdapat beberapa kekurangan yang dapat dikembangkan pada penelitian selanjutnya. Oleh karena itu untuk pengembangan selanjutnya disarankan:

a. Dalam menganalisis data *twitter* bisa menggunakan metode dari teknik data mining lainnya sebagai pembanding.
b. Untuk membandingkan ada atau tidaknya perbedaan mengenai topik yang sering dikicaukan antara dua orang atau lebih, dapat menggunakan *comparison wordcloud*.
c. Menambah parameter pada data *tweet* yang dianalisis, seperti *in replay to status id, latitude, longitude, place id, display coordinates, trims user* atau *includes entities*.

DAFTAR PUSTAKA

Abimanyu, Aditya. 2012. *Analisa Media Sosial Twitter dengan Perhitungan Graph Edit Distance untuk Mendeteksi Rumor pada Trending Topic SIAK-NG*. https://www.lib.ui.ac.id 10 Desember 2016.

Adi, Wibawa Putra. 2013. *Media Sosial dan Jejaring Sosial*. https://wibawaadiputra.wordpress.com/ 13 Desember 2016

Adiyana, Imam. 2015. *Implementasi Text Mining Pada Mesin Pencarian Twitter Untuk Menganalisis Topik–Topik Terkait "Kpk Dan Jokowi"*. https://publikasiilmiah.ums.ac.id/ 12 Desember 2016

Anas, Muhammad Faishol. 2011. *Implementasi Text Mining Untuk Mendukung Pencarian Topik Pada E-Library Menggunakan Mobile Device*. https://www.undana.ac.id 10 Desember 2016

Anonim. *Social Media Trends: A Few Interesting Developments*. http://tnmedia.com/social-media-trends-interesting-developments/. 13 Januari 2017

Dahlan, Ahmad. 2015. *Definisi Sampling Serta Jenis Metode dan Teknik Sampling*. http://www.eurekapendidikan.com/ 11 Desember 2016

Kestrilia dan Hendra. Juni 2014. *Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering*. Vol 2, No 1. https://cybermatika.stei.itb.ac.id/ 10 Desember 2016

Kurniawan, Bambang. 2012. *Klasifikasi Konten Berita dengan Metode Text Mining*. Vol 1 No 1. http://jurnal.usu.ac.id/ 10 Desember 2016

Nhuri, Riri. 2009. *Biografi Barack Obama*. http://www.biografiku.com/ 11 Desember 2016

_____. 2016. *Biografi Donald Trump*. http://www.biografiku.com/ 11 Desember 2016

Retnawiyati, Eka. 2015. *Analisis Sentimen Pada Data Twitter dengan Menggunakan Text Mining terhadap Suatu Produk*. https://www.if.binadarma.ac.id/ 10 Desember 2016

Rizal, Muhammad Fikri. 2016. *Inilah 10 Pria Paling Berpengaruh di Twitter*. http://www.solopos.com/ 13 Desember 2016

Sanjaya, Suwanto. 2015. *Pengelompokan Dokumen Menggunakan Winnowing Fingerprint dengan Metode K-Nearest Neighbour*. Vol 1, No 2. https://ejournal.uin-suska.ac.id/ 10 Desember 2016

Sano, Dian. 2015. *Proses dalam Text Mining - Seri Text Mining dan Web Mining*. https://beritati.blogspot.co.id/2015/05/proses-dalam-text-mining-seri-text.html 1 Maret 2017

Sasrawan, Hedi. 2013. *Pengertian Sosialisasi*. http://hedisasrawan.blogspot.com/ 11 Desember 2016

Supriyaningsih., Maulina. 2013. *Clouds of The Gates.* https://www.academia.edu 12 Desember 2016

Lampiran 1. Script yang digunakan untuk mengautentikasi TwitterAPI

```
#Inisialisasivariabeltwitterapi
cons_key <- '<consumer_key>'
cons_sec <- '<consumer_secret>'
acc_token <- '<access_token>'
acc_sec <- '<access_secret>'

#autentikasiTwitterAPI
library(twitteR)
setup_twitter_oauth(
 consumer_key="geCwdIgMD9QkM2rdyB422ZuYJ",
consumer_secret="VU0qrzcYNEdd6C5E4ql3JV8W9djk053kHUm51MwBQrpUNAKrEx",
 access_token="2360885588-kYE5NIK0561CIGqTCNyVgHeSlVgRzBWAN906JFF",
 access_secret="H6OcgVypfASFPQsfGnAFOoRCRGAyhn0wY0DL5bQmWJnR2"
)
```

Lampiran 2. Script yang digunakan untuk melakukan analisis text mining pada akun Twitter @BarackObama

```
tweetsa=userTimeline('BarackObama',n=2000,maxID=NULL,sinceID=NULL,include
Rts=FALSE,excludeReplies=FALSE)
text = sapply(tweetsa, function(x) x$getText())

tweets.df<-twListToDF(tweetsa)
dim(tweets.df)
for(i in c(1:15,2000)) {
cat(paste0("[", i,"] "))
writeLines(strwrap(tweets.df$text[i],60))
}

library(tm)
# build a corpus, and specify the source to be character vectors
myCorpus<-Corpus(VectorSource(tweets.df$text))

# hanya mengambil huruf dan spasi
removeNumPunct<-function(x)gsub("[^[:alpha:][:space:]]*","", x)
myCorpus<-tm_map(myCorpus,content_transformer(removeNumPunct))

# remove URLs
removeURL<-function(x)gsub("http[^[:space:]]*","", x)
myCorpus<-tm_map(myCorpus,content_transformer(removeURL))

# remove punctuation
myCorpus <- tm_map(myCorpus, removePunctuation)

# remove numbers
myCorpus <- tm_map(myCorpus, removeNumbers)
```

```
# remove URLs
removeURL<-function(x)gsub("http[^[:space:]]*","", x)
myCorpus<-tm_map(myCorpus,content_transformer(removeURL))

# convert to lower case
myCorpus<-tm_map(myCorpus,content_transformer(tolower))

# add some extra stop words: "dan","itu", etc.
myStopwords<-c(stopwords('english'),"rt","I","and","in","the",
"be","can","or","who","was","is","to","for","oh","my","on","of","a")

# remove stopwords from corpus
myCorpus<-tm_map(myCorpus, removeWords, myStopwords)

# remove extra whitespace
myCorpus<-tm_map(myCorpus, stripWhitespace)
myCorpusCopy<-myCorpus

# stem words
myCorpus<-tm_map(myCorpus,stemDocument)
for(i in c(1:15,2000)) {
cat(paste0("[", i,"] "))
writeLines(strwrap(as.character(myCorpus[[i]]),60))
}

stemCompletion2<-function(x,dictionary) {
x<-unlist(strsplit(as.character(x)," "))
x<-x[x!=""]
x<-stemCompletion(x,dictionary=dictionary)
x<-paste(x,sep="",collapse=" ")
PlainTextDocument(stripWhitespace(x))
}
```

```
myCorpus<-lapply(myCorpus, stemCompletion2,dictionary=myCorpusCopy)
myCorpus<-Corpus(VectorSource(myCorpus))

miningCases<-lapply(myCorpusCopy,
function(x) { grep(as.character(x),pattern=" \\<mining")} )
sum(unlist(miningCases))
minerCases<-lapply(myCorpusCopy,
function(x) {grep(as.character(x),pattern=" \\<miner")} )
sum(unlist(minerCases))

tdm<- TermDocumentMatrix(myCorpus,
control=list(wordLengths=c(1,Inf)))
tdm

idx<-which(dimnames(tdm)$Terms=="BarackObama")
inspect(tdm[idx+(0:3),100:110])

findFreqTerms(tdm,lowfreq=30)
term.freq<-rowSums(as.matrix(tdm))
term.freq<-subset(term.freq,term.freq>=30)
df<-data.frame(term=names(term.freq),freq= term.freq)

library(ggplot2)
ggplot(df,aes(x= term,y= freq))+geom_bar(stat="identity")+
xlab("Terms")+ylab("Count")+coord_flip()
barplot(term.freq, las=2) # barplot kata

#Frequent Terms and Associations (kata yang sering keluar dan asosiasi kata)
findFreqTerms(tdm, lowfreq=30)
findAssocs(tdm, 'senate', 0.2)
findAssocs(tdm, 'obama', 0.2)
findAssocs(tdm, 'president', 0.2)
```

```
# membuat awan kata (word cloud)
library(wordcloud)
m<-as.matrix(tdm)
wordFreq <- sort(rowSums(m), decreasing=TRUE)
set.seed(375)

pal<-brewer.pal(10,"Set3")[-(1:2)]
wordcloud(words=names(wordFreq), freq=wordFreq, min.freq=15, random.order=F,
colors=pal)

#Clustering Words
# remove sparse terms
tdm2 <- removeSparseTerms(tdm, sparse=0.95)
m2 <- as.matrix(tdm2)
# cluster terms
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method="ward.D")
plot(fit)
# cut tree into 5 clusters
rect.hclust(fit, k=5)
(groups <- cutree(fit, k=5))

#Clustering Tweets #Clustering Tweets with the k-means Algorithm
#transpose the matrix to cluster documents (tweets)
m3 <- t(m2)
set.seed(122) # set a fixed random seed

# k-means clustering of tweets
k <- 5
kmeansResult <- kmeans(m3, k)
# cluster centers
round(kmeansResult$centers, digits=3)
for (i in 1:k) {
cat(paste("cluster ", i, ": ", sep=""))
s <- sort(kmeansResult$centers[i,], decreasing=T)
cat(names(s)[1:3], "\n")
}
```

Lampiran 3. Script yang digunakan untuk melakukan analisis text mining pada akun Twitter @realDonaldTrump

```
tweets.df<-twListToDF(tweetsa)
dim(tweets.df)
for(i in c(1:15,2000)) {
cat(paste0("[", i,"] "))
writeLines(strwrap(tweets.df$text[i],60))
}

library(tm)
# build a corpus, and specify the source to be character vectors
myCorpus<-Corpus(VectorSource(tweets.df$text))

# hanya mengambil huruf dan spasi
removeNumPunct<-function(x)gsub("[^[:alpha:][:space:]]*","", x)
myCorpus<-tm_map(myCorpus,content_transformer(removeNumPunct))

# remove URLs
removeURL<-function(x)gsub("http[^[:space:]]*","", x)
myCorpus<-tm_map(myCorpus,content_transformer(removeURL))

# remove punctuation
myCorpus <- tm_map(myCorpus, removePunctuation)
# remove numbers
myCorpus <- tm_map(myCorpus, removeNumbers)
```

```
# remove URLs
removeURL<-function(x)gsub("http[^[:space:]]*","", x)
myCorpus<-tm_map(myCorpus,content_transformer(removeURL))

# convert to lower case
myCorpus<-tm_map(myCorpus,content_transformer(tolower))

# add some extra stop words: "dan","itu", etc.
myStopwords<-
c(stopwords('english'),"rt","I","and","in","the","be","can","or","who","was","is"))

# remove stopwords from corpus
myCorpus<-tm_map(myCorpus, removeWords, myStopwords)

# remove extra whitespace
myCorpus<-tm_map(myCorpus, stripWhitespace)
myCorpusCopy<-myCorpus

# stem words
myCorpus<-tm_map(myCorpus,stemDocument)
for(i in c(1:15,2000)) {
cat(paste0("[", i,"] "))
writeLines(strwrap(as.character(myCorpus[[i]]),60))
}

stemCompletion2<-function(x,dictionary) {
x<-unlist(strsplit(as.character(x)," "))
x<-x[x!=""]
x<-stemCompletion(x,dictionary=dictionary)
x<-paste(x,sep="",collapse=" ")
PlainTextDocument(stripWhitespace(x))
}
```

```
myCorpus<-lapply(myCorpus, stemCompletion2,dictionary=myCorpusCopy)
myCorpus<-Corpus(VectorSource(myCorpus))

miningCases<-lapply(myCorpusCopy,
function(x) { grep(as.character(x),pattern=" \\<mining")} )
sum(unlist(miningCases))
minerCases<-lapply(myCorpusCopy,
function(x) {grep(as.character(x),pattern=" \\<miner")} )
sum(unlist(minerCases))

tdm<- TermDocumentMatrix(myCorpus,
control=list(wordLengths=c(1,Inf)))
tdm

idx<-which(dimnames(tdm)$Terms=="realDonaldTrump")
inspect(tdm[idx+(0:3),100:110])

findFreqTerms(tdm,lowfreq=30)
term.freq<-rowSums(as.matrix(tdm))
term.freq<-subset(term.freq,term.freq>=30)
df<-data.frame(term=names(term.freq),freq= term.freq)

library(ggplot2)
ggplot(df,aes(x= term,y= freq))+geom_bar(stat="identity")+
xlab("Terms")+ylab("Count")+coord_flip()
barplot(term.freq, las=2) # barplot kata

#Frequent Terms and Associations (kata yang sering keluar dan asosiasi kata)
findFreqTerms(tdm, lowfreq=30)
findAssocs(tdm, will', 0.2)
findAssocs(tdm, 'great', 0.2)
```

```
# membuat awan kata (word cloud)
library(wordcloud)
m<-as.matrix(tdm)
wordFreq <- sort(rowSums(m), decreasing=TRUE)
set.seed(375)

pal<-brewer.pal(10,"Set3")[-(1:2)]
wordcloud(words=names(wordFreq), freq=wordFreq, min.freq=15, random.order=F,
colors=pal)

#Clustering Words
# remove sparse terms
tdm2 <- removeSparseTerms(tdm, sparse=0.95)
m2 <- as.matrix(tdm2)
# cluster terms
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method="ward.D")
plot(fit)
# cut tree into 5 clusters
rect.hclust(fit, k=5)
(groups <- cutree(fit, k=5))

#Clustering Tweets #Clustering Tweets with the k-means Algorithm
#transpose the matrix to cluster documents (tweets)
m3 <- t(m2)
set.seed(122) # set a fixed random seed

# k-means clustering of tweets
k <- 5
kmeansResult <- kmeans(m3, k)
# cluster centers
round(kmeansResult$centers, digits=3)
for (i in 1:k) {
cat(paste("cluster ", i, ": ", sep=""))
s <- sort(kmeansResult$centers[i,], decreasing=T)
cat(names(s)[1:3], "\n")
}
```

Lampiran 4. Script yang digunakan untuk melakukan analisis sentimen pada akun Twitter @realDonaldTrump

```
library(twitteR)
library(RCurl)
library(RJSONIO)
library(stringr)
library(tm)
library(wordcloud)
#############################################################################
getSentiment <- function (text, key){

  text <- URLencode(text);

  #save all the spaces, then get rid of the weird characters that break the API, then
convert back the URL-encoded spaces.
  text <- str_replace_all(text, "%20", " ");
  text <- str_replace_all(text, "%\\d\\d", "");
  text <- str_replace_all(text, " ", "%20");


  if (str_length(text) > 360){
    text <- substr(text, 0, 359);
  }
  ##########################################

  data <-
getURL(paste("http://api.datumbox.com/1.0/TwitterSentimentAnalysis.json?api_key=
", "8690e804e964442f5751ced6e64825c2", "&text=",text, sep=""))

  js <- fromJSON(data, asText=TRUE);

  # get mood probability
  sentiment = js$output$result

  ###################################

    return(list(sentiment=sentiment))
}

clean.text <- function(some_txt)
{
  some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", some_txt)
  some_txt = gsub("@\\w+", "", some_txt)
  some_txt = gsub("[[:punct:]]", "", some_txt)
  some_txt = gsub("[[:digit:]]", "", some_txt)
  some_txt = gsub("http\\w+", "", some_txt)
  some_txt = gsub("[ \t]{2,}", "", some_txt)
```

```
  some_txt = gsub("^\\s+|\\s+$", "", some_txt)
  some_txt = gsub("amp", "", some_txt)
  # define "tolower error handling" function
  try.tolower = function(x)
  {
    y = NA
    try_error = tryCatch(tolower(x), error=function(e) e)
    if (!inherits(try_error, "error"))
      y = tolower(x)
    return(y)
  }

  some_txt = sapply(some_txt, try.tolower)
  some_txt = some_txt[some_txt != ""]
  names(some_txt) = NULL
  return(some_txt)
}

#################################################################

print("Getting tweets...")
# get some tweets
tweets =
serTimeline('BarackObama',n=2000,maxID=NULL,sinceID=NULL,includeRts=FAL
SE,excludeReplies=FALSE)
# get text
tweet_txt = sapply(tweets, function(x) x$getText())

# clean text
tweet_clean = clean.text(tweet_txt)
tweet_num = length(tweet_clean)
# data frame (text, sentiment)
tweet_df = data.frame(text=tweet_clean, sentiment=rep("",
tweet_num),stringsAsFactors=FALSE)

print("Getting sentiments...")
# apply function getSentiment
sentiment = rep(0, tweet_num)
for (i in 1:tweet_num)
{
  tmp = getSentiment(tweet_clean[i], db_key)
    tweet_df$sentiment[i] = tmp$sentiment
    print(paste(i," of ", tweet_num))
    }
```

```
 # delete rows with no sentiment
tweet_df <- tweet_df[tweet_df$sentiment!="",]

#separate text by sentiment
sents = levels(factor(tweet_df$sentiment))
#emos_label <- emos

# get the labels and percents

labels <-  lapply(sents, function(x)
paste(x,format(round((length((tweet_df[tweet_df$sentiment
==x,])$text)/length(tweet_df$sentiment)*100),2),nsmall=2),"%"))

nemo = length(sents)
emo.docs = rep("", nemo)
for (i in 1:nemo)
{
  tmp = tweet_df[tweet_df$sentiment == sents[i],]$text

  emo.docs[i] = paste(tmp,collapse=" ")
}

# remove stopwords
emo.docs = removeWords(emo.docs, stopwords("german"))
emo.docs = removeWords(emo.docs, stopwords("english"))
corpus = Corpus(VectorSource(emo.docs))
tdm = TermDocumentMatrix(corpus)
tdm = as.matrix(tdm)
colnames(tdm) = labels

# comparison word cloud
comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
                 scale = c(3,.5), random.order = FALSE, title.size = 1.5)
```

Lampiran 5. Script yang digunakan untuk melakukan analisis sentimen pada akun Twitter @realDonaldTrump

```
library(twitteR)
library(RCurl)
library(RJSONIO)
library(stringr)
library(tm)
library(wordcloud)
#############################################################################
getSentiment <- function (text, key){

  text <- URLencode(text);

  #save all the spaces, then get rid of the weird characters that break the API, then
convert back the URL-encoded spaces.
  text <- str_replace_all(text, "%20", " ");
  text <- str_replace_all(text, "%\\d\\d", "");
  text <- str_replace_all(text, " ", "%20");


  if (str_length(text) > 360){
    text <- substr(text, 0, 359);
  }
  ##########################################

  data <-
getURL(paste("http://api.datumbox.com/1.0/TwitterSentimentAnalysis.json?api_key=
", "8690e804e964442f5751ced6e64825c2", "&text=",text, sep=""))

  js <- fromJSON(data, asText=TRUE);

  # get mood probability
  sentiment = js$output$result

  ###################################

    return(list(sentiment=sentiment))
}

clean.text <- function(some_txt)
{
  some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", some_txt)
  some_txt = gsub("@\\w+", "", some_txt)
  some_txt = gsub("[[:punct:]]", "", some_txt)
  some_txt = gsub("[[:digit:]]", "", some_txt)
  some_txt = gsub("http\\w+", "", some_txt)
  some_txt = gsub("[ \t]{2,}", "", some_txt)
```

```
  some_txt = gsub("^\\s+|\\s+$", "", some_txt)
  some_txt = gsub("amp", "", some_txt)
  # define "tolower error handling" function
  try.tolower = function(x)
  {
    y = NA
    try_error = tryCatch(tolower(x), error=function(e) e)
    if (!inherits(try_error, "error"))
      y = tolower(x)
    return(y)
  }

  some_txt = sapply(some_txt, try.tolower)
  some_txt = some_txt[some_txt != ""]
  names(some_txt) = NULL
  return(some_txt)
}

#################################################################

print("Getting tweets...")
# get some tweets
tweets = userTimeline('realDonaldTrump', n=2000,maxID=NULL,sinceID=NULL,
includeRts=FALSE,excludeReplies=FALSE)
# get text
tweet_txt = sapply(tweets, function(x) x$getText())

# clean text
tweet_clean = clean.text(tweet_txt)
tweet_num = length(tweet_clean)
# data frame (text, sentiment)
tweet_df = data.frame(text=tweet_clean, sentiment=rep("",
tweet_num),stringsAsFactors=FALSE)

print("Getting sentiments...")
# apply function getSentiment
sentiment = rep(0, tweet_num)
for (i in 1:tweet_num)
{
  tmp = getSentiment(tweet_clean[i], db_key)
    tweet_df$sentiment[i] = tmp$sentiment
    print(paste(i," of ", tweet_num))
    }
```

```
 # delete rows with no sentiment
tweet_df <- tweet_df[tweet_df$sentiment!="",]

#separate text by sentiment
sents = levels(factor(tweet_df$sentiment))
#emos_label <- emos

# get the labels and percents

labels <-  lapply(sents, function(x)
paste(x,format(round((length((tweet_df[tweet_df$sentiment
==x,])$text)/length(tweet_df$sentiment)*100),2),nsmall=2),"%"))

nemo = length(sents)
emo.docs = rep("", nemo)
for (i in 1:nemo)
{
  tmp = tweet_df[tweet_df$sentiment == sents[i],]$text

  emo.docs[i] = paste(tmp,collapse=" ")
}

# remove stopwords
emo.docs = removeWords(emo.docs, stopwords("german"))
emo.docs = removeWords(emo.docs, stopwords("english"))
corpus = Corpus(VectorSource(emo.docs))
tdm = TermDocumentMatrix(corpus)
tdm = as.matrix(tdm)
colnames(tdm) = labels

# comparison word cloud
comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
                 scale = c(3,.5), random.order = FALSE, title.size = 1.5)
```