Text Mining pada Media Sosial Twitter Studi Khasus: Pilkada DKI 2017 Putaran 2

HASIL

Oleh:
Dimas Bagus C. W.
NIM 151820101014

PROGRAM STUDI PASCASARJANA MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS JEMBER
2017

Text Mining pada Media Sosial Twitter
Studi Kasus: Masa Tenang Pilkada DKI 2017 Putaran 2

TESIS

diajukan guna melengkapi tugas akhir dan memenuhi salah satu syarat
untuk menyelesaikan Program Studi Magister Matematika (S2)
serta mencapai gelar Magister Sains

Oleh:
Dimas Bagus C. W.
NIM 151820101014

PROGRAM STUDI PASCASARJANA MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS JEMBER
2017

PERSEMBAHAN

Tesis ini saya persembahkan untuk:

1. Papah Suarji dan Mamah P. W. Handayani serta adek Rara Diyah Ayu Candra Diana yang senantiasa memberi doa, semangat, dan kasih sayang;
2. Bapak Alfian Futuhul Hadi yang selalu mendorong dan memotivasi dalam menentukan langkah-langkah di perguruan tinggi ini;
3. Semua guru dan dosen yang telah memberikan ilmu serta dukungan dalam setiap tahap-tahap pada proses pertumbuhan di lingkungan akademik maupun non akademik;

MOTTO

“demi masa, sungguh manusia berada dalam kerugian,
kecuali orang-orang yang beriman dan mengerjakan kebajikan
serta saling menasehati untuk kebenaran
dan saling menasehati untuk kesabaran”
(Al-‘Asr : ayat 1-3)

PERNYATAAN

Saya yang bertandatangan di bawah ini:

Nama : Dimas Bagus C. W.

NIM : 151820101014

Menyatakan dengan sesungguhnya bahwa karya ilmiah yang berjudul "*Text Mining* pada Media Sosial Twitter Studi Kasus: Masa Tenang Pilkada DKI 2017 Putaran 2" adalah benar- benar hasil karya sendiri, kecuali kutipan yang sudah saya sebutkan sumbernya, belum pernah diajukan dalam institusi manapun, dan bukan karya jiplakan. Saya bertanggung jawab atas keabsahan dan kebenaran isinya sesuai dengan sikap ilmiah yang harus dijunjung tinggi. Demikian pernyataan ini saya buat dengan sebenarnya, tanpa ada tekanan dan paksaan dari pihak manapun serta bersedia mendapat sanksi akademik jika ternyata di kemudian hari pernyataan ini tidak benar.

Jember, 17 Juli 2017

Yang menyatakan,

Dimas Bagus C. W.
NIM 151820101014

Text Mining pada Media Sosial Twitter
Studi Kasus: Masa Tenang Pilkada DKI 2017 Putaran 2

Oleh :
Dimas Bagus C. W.
NIM 151820101014

Pembimbing

Dosen Pembimbing Utama	:	Dr. Alfian Futuhul Hadi, S.Si., M. Si.
Dosen Pembimbing Anggota	:	Drs. Moh. Hasan, M.Sc., Ph.D.

PENGESAHAN

Tesis berjudul "*Text Mining* pada Media Sosial Twitter Studi Kasus: Masa Tenang Pilkada DKI 2017 Putaran 2" telah diuji dan disahkan pada:

Hari, tanggal :

Tempat : Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Jember

Tim penguji

Dosen Pembimbing Utama,	Dosen Pembimbing Anggota,
Dr. Alfian Futuhul Hadi, S.Si., M. Si. NIP. 197407192000121001	Drs. Moh. Hasan, M.Sc., Ph.D. NIP. 196404041988021001

Anggota tim penguji

Dosen Penguji I,	Dosen Penguji II,
Drs. Budi Lestari, PGD.Sc., M.Si. NIP. 196310251991031003	Dian Anggraeni, S.Si., M.Si. NIP. 198202162006042002

Mengesahkan,
Dekan

Drs. Sujito, Ph.D.
NIP 196102041987111001

RINGKASAN

***Text Mining* pada Media Sosial Twitter Studi Kasus: Pilkada DKI 2017 Putaran 2**; Dimas Bagus C. W.; 2017; 74 halaman; Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Jember.

Penelitian ini bertujuan untuk menggunakan metode *text mining* dalam mengolah data teks pada media sosial Twitter untuk dapat menemukan suatu informasi, baik berupa ringkasan dari *tweets* hingga menggunakannya untuk menduga hasil dari pencoblosan. Peneliti mengumpulkan data *tweets* dimana 20.000 data *tweets* berhubungan dengan kata "anies" dan 20.000 data *tweets* berhubungan dengan kata "ahok" yang di ambil setiap hari selama tanggal 15-19 April 2017 (masa tenang hingga waktu pencoblosan). Teknik analisis data yang dilakukan adalah model alir dengan tahap-tahap sebagai berikut; (a) *text preprocessing* (b) pembobotan (c) *unsupervised learning* dan *supervised learning* (d) pembahasan.

Teknik analisis diawali dengan text preprocessing. Tahap ini bertujuan untuk membersihkan serta mereduksi data-data yang tidak diperlukan seperti simbol, tanda baca, alamat *link*, dan *stopwords*. Kemudian peneliti memberikan bobot pada masing-masing data *tweets* dengan menghitung besar *Term Frequency – Inverse Term Frequence* (TF-IDF). Setiap data *tweets* yang memiliki besar TF-TDF yang rendah tidak diikutkan dalam proses selanjutnya. Dengan menggunakan metode penentuan sentimen serta Naive Bayes pada *supervised learing* dan metode pengelompokkan dengan K-Means seerta *Topic Modeling* pada *unsupervised learning*, peneliti mencari pola yang terbentuk serta hal menarik lainnya yang dapat menjadi pembahasan. Selanjutnya dengan memvisualisasikannya (berdasarkan sentimen ataupun kelompok yang terbentuk), dapat memberikan pembahasan yang lebih menarik.

Penentuan sentimen pada masing-masing *tweets* dilakukan dengan membandingkan banyak kata-kata yang bersentimen negatif dan kata-kata yang bersentimen positif. Apabila dalam satu *tweets* memiliki jumlah kata yang bersentimen negatif lebih banyak dari pada kata yang bersentimen positif maka *tweets* tersebut bersentimen negatif. Begitu pula sebaliknya apabila jumlah kata bersentimen positif lebih banyak. Selain itu, *tweets* bersentimen netral. Sedangkan metode Naive Bayes digunakan untuk melakukan prediksi terhadap data baru. Penelitian ini menggunakan metode ini untuk dapat mengetahui sejauh mana metode ini dapat melakukan prediksi sentimen pada data tersebut.

Metode-metode pengelompokkan bertujuan untuk mengumpulkan data berdasarkan karakteristik ataupun topik yang muncul. Dengan memperoleh kelompok-kelompok yang baru (5, 10, 25, 50, 75, dan 100 kelompok) , peneliti akan lebih mudah dalam mengambil kesimpulan ataupun ringkasan pada data *tweets* yang sangat banyak tersebut.

Hasil penelitian ini menunjukan bahwa terdapat kata-kata khusus yang menggambarkan sentimen pada *tweets* tersebut. Selain itu *hastag* yang digunakan oleh user terhadap pasangan calon juga menjadi pendukung sentimen yang muncul

pada *tweets* tersebut. Oleh karennya, peneliti menghimpun *hastag* dan kata-kata yang memiliki sentimen khusus tersebut yang selanjutnya digunakan untuk memberikan sentimen pada data *tweets*.

Metode Naive Bayes hanya mampu memberikan prediksi sentimen yang baik pada saat digunakan memprediksi data *tweets* yang diambil pada hari yang sama. Apabila data hari ini digunakan untuk melakukan prediksi pada esok hari ataupun hari yang lain, metode ini hanya dapat memberikan keakuratan prediksi yang kurang dari 50%. Hal ini disebabkan oleh keragaman data *tweets* pada hari ini tidak sama dengan *tweets* pada hari esok ataupun hari yang lain. Sehingga probabilitas dari masing-masing kata menjadi tidak sama.

Pada proses pengelompokkan, metode K-Means memberikan hasil kelompok yang tidak merata. Selalu ada 1 kelompok pada K-Means memiliki anggota (*tweets*) yang lebih banyak dari kelompok yang lain. Berbeda dengan K-Means, pengolompokkan berdasarkan topik yang terbentuk yang dilakukan oleh Topic Modeling dapat memberikan hasil pengelompokkan yang lebih merata. Hal ini terjadi karena proses pengelompokkan K-Means terlalu sederhana yaitu dengan menghitung jarak terdekat pada masing-masing titik (bobot TF-TDF). Topic Modeling bekerja dengan melakukan iterasi terhadap masing-masing kata dan masing-masing topik yang terbentuk. Sehingga hal ini memberikan hasil pengelompokkan yang lebih merata dan sesuai dengan datanya.

Hal lain yang menarik perhatian adalah pola sentimen yang muncul pada data *tweets* “anies” dan data *tweets* “ahok”. Pada hari pertama masa tenang (16 April 2017), kedua data memiliki jumlah sentimen netral, positif dan negatif yang hampir sama. Namun pada keesokan harinya (17 April 2017), setiap data *tweets* baik “anies” dan “ahok” mendapatkan sentimen negatif yang lebih banyak dari pada hari sebelumnya. Bahkan data *tweets* “ahok” mendapatkan lebih dari 12.000 data yang bersentimen negatif dari *user* sedangkan data *tweets* “anies” mencapai 8.000 data. Pada hari selanjutnya (18 April 2017), tren yang terjadi adalah *tweets* bersentimen positif mendominasi data *tweets* pada masing-masing pasangan calon. Namun jumlah *tweets* bersentimen positif dari data “ahok” lebih sedikit dari jumlah *tweets* bersentimen positif dari data “anies”.

Pola sentimen yang muncul pada 3 hari tersebut (16-18 April 2017) memberikan indikasi bahwa *user* pada media Twitter lebih mendukung pasangan Anies-Sandi daripada pasangan Ahok-Djarot. Hal ini ternyata sejalan dengan hasil perhitungan suara yang dilakukan oleh KPU yang menyatakan bahwa pasangan Anies-Sandi menang dalam pemilihan Kepala Daerah DKI Jakarta Putaran 2.

PRAKATA

Puji syukur kehadirat Allah SWT. yang tak henti-hentinya melimpahkan rahmat, memberikan tuntunan, taufik, karunia, hidayah, dan inayah Nya sehingga tesis yang berjudul "*Text Mining* pada Media Sosial Twitter Studi Kasus: Pilkada DKI Putaran 2" dapat terselesaikan sesuai dengan waktu yang telah ditentukan oleh Nya. Tesis ini disusun untuk memenuhi salah satu syarat dalam menyelesaikan program Magister di Jurusan Matematika Fakultas Matematika dan Ilmu Penegetahuan Alam Universitas Jember. Sholawat serta salam semoga selalu tercurahkan keharibaan beliau junjungan kami Nabi Muhammad SAW yang telah menjadi rahmatan lil'alamin.

Penyusunan tesis ini tidak terlepas dari bantuan berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. Drs. Sujito, Ph. D. dan Kusbudiono, S.Si., M.Si. selaku Dekan dan Ketua Jurusan Matematika Fakultas MIPA Universitas Jember yang telah memberikan fasilitas-fasilitas dalam tahap perkuliahan;
2. Dr. Alfian Futuhul Hadi, S.Si., M.Si. selaku Dosen Pembimbing Utama sekaligus Dosen Pembimbing Akademik dan Drs. Moh. Hasan, M.Sc., Ph.D. selaku Dosen Pembimbing Anggota yang telah memberikan bimbingan dan bantuan untuk penyempurnaan tesis ini;
3. Drs. Budi Lesatari, PGD.Sc., M.Si. selaku Dosen Penguji I dan Dian Anggraeni, S.Si, M.Si. selaku Dosen Penguji II yang telah memberikan kritik dan saran yang membangun untuk penyempurnaan tesis ini;
4. Seluruh dosen dan karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam yang telah memberikan ilmu serta membantu selama proses perkuliahan berlangsung;
5. Papah Suarji dan Mamah P. W. Handayani serta adek tercinta Rara Diyah Ayu Candra Diana yang senantiasa memberi doa, semangat, dan kasih sayang;

6. Sahabat/i PMII FMIPA Universitas Jember;
7. Keluarga Masjid Sunan Kalijaga yang selalu memberikan dukungan dalam hal apapun.

Penulis menyadari bahwa dalam menyusun tesis ini masih terdapat kekurangan baik isi maupun susunannya. Oleh karena itu, penulis mengharapkan saran dan kritik demi penyempurnaan tesis ini. Akhirnya penulis berharap semoga tesis ini dapat memberi manfaat dan sumbangan bagi pembaca.

Jember, 17 Juli 2017

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1. PENDAHULUAN

1.1 Latar Belakang

Sekitar ber-*Terabyte* atau ber-*Petabyte* dari data tertuang pada jaringan komputer, *World Wide Web* (WWW), dan berbagai *platform* penyimpanan data setiap harinya, dari bisnis, sosial, sains dan teknik, pengobatan, dan hampir setiap aspek lain dari kehidupan sehari-hari (Han, 2012). Sebagian besar data tersebut tertuang kedalam bentuk data teks, baik yang terstruktur maupun tidak tersturktur. Namun jarang sekali data-data yang berupa teks tersebut digunakan dalam melakukan penelitian. Padahal banyak informasi yang dapat diperoleh dari mengolah data-data teks tersebut terutama dalam mengetahui opini masyarakat terhadap suatu produk tertentu.

Seiring dengan perkembangan jaman, teknik-teknik pengolahan data yang berkapasitas besar seperti data penjualan, teks dan perbankan mulai bermunculan dan semakin berkembang. Perkembangan pengolahan data yang sangat banyak tersebut kemudian mengerucut pada satu bidang keilmuan yaitu *data mining*. *Data mining* (penggalian data) merupakan proses untuk menjelajahi pola dan pengetahuan tertentu dari data yang sangat banyak seperti data yang terdapat pada jaringan komputer. *Data mining* bekerja dengan beberapa tahap yaitu, pembersihan data yang tidak perlu (*cleaning*), keintegrasian antar data (*integration*), seleksi (*selection*), transformasi (*transformation*), setelah itu barulah digali (*mining*) setiap informasi yang telah diproses tadi. Proses pengolahan data yang sangat banyak dengan metode penggalian data juga berkembang ke arah pengolahan data teks, yang dikenal dengan istilah *text mining*.

Text mining berfokus terhadap data teks tidak terstruktur seperti opini publik, artikel, buku, maupun berita *online*. *Text mining* bekerja dengan melakukan pembelajaran pada data. Pembelajaran yang dilakukan tersebut terbagi menjadi 3 jenis bergantung pada jenis data teks yang digunakan. Pertama adalah *Supervised Learning*, yaitu pembelajaran pada data teks yang telah memiliki label atau data yang

telah terklasifikasi sebelumnya. Hasil pembelajaran ini akan digunakan untuk memberikan label ataupun klasifikasi pada data lain yang masing belum memiliki label. Kedua adalah *Unsupervised Learning*, yaitu pembelajaran pada data yang masih belum memiliki klasifikasi ataupun label. Pembelajaran ini umumnya digunakan untuk mempelajari karakteristik maupun topik yang muncul pada data dan mengkelompokan berdasarkan karakteristik serta topik-topik tersebut. Kemudian yang ketiga adalah *Semi Unsupervised Learning*, yaitu pembelajaran yang digunakan pada data teks campuran (ada yang memiliki label, ada pula yang tidak). Pembelajaran yang dilakukan kedua metode ini menggunakan kombinasi dari *Supervised Learning* dan *Unsupervised Learning*.

Pembelajaran pada jenis data yang telah memiliki label (*Supervised Learning*) memungkinkan penggunanya ini untuk memprediksi label yang muncul pada suatu kumpulan data teks yang masih belum memiliki label. Untuk dapat memprediksi label pada data teks, diperlukan salah satu metode pembelajaran *Supervised* yang mampu melakukan pembelajaran pada kumpulan data teks tersebut. Salah satu diantaranya adalah metode Naive Bayes. Metode ini melakukan pembelajaran terhadap teks dengan mempelajari seberapa besar probabilitas suatu kata masuk kedalam suatu label. Sehingga untuk dapat menduga label yang muncul, maka metode ini memerlukan suatu data set terlebih dahulu yang di kenal dengan istilah *data training*. Hasil dari pembelajaran pada *data training* tersebut kemudian digunakan untuk mempelajari label pada data baru atau *data test*.

Selain itu, *text mining* juga dapat digunakan untuk melakukan eksplorasi data teks dengan menemukan hubungan yang terjadi antar kata berdasarkan besar nilai *Confidence* dan *Lift* kata-kata tersebut dengan kata-kata yang lain. Metode ini dikenal dengan istilah *Association Rules*. Disisi lain, metode-metode *text mining* lainnya berfokus terhadap pengelompokan data yang masih belum memiliki label (*Unsupervised*), baik pengelompokan berdasarkan karakteristiknya ataupun dengan berdasarkan topik pembahasannya. Metode-metode pengelompokan tersebut diantaranya adalah K-Means dan *Topic Modeling*. Jenis data *Unsupervised* lebih

sering ditemui pada sistem jaringan komputer manusia. Sehingga dalam perkembangannya metode ini lebih sering digunakan dan mengalami perkembangan-perkembangan dalam pengaplikasiannya.

Pilkada DKI adalah salah satu pemilihan kepala daerah yang selalu menjadi pembicaraan publik. Hal ini disebabkan oleh karena DKI merupakan ibu kota negara sekaligus menjadi jantung perpolitikan di Indonesia. Maka tak mengherankan apabila DKI menjadi contoh untuk daerah-daerah lain. Sorotan pada proses Pilkada DKI juga terjadi di berbagai macam media baik cetak, elektronik hingga media sosial. Media sosial menjadi salah satu media yang efektif bagi pasangan calon maupun tim suksesnya untuk melakukan kampanye ataupun memunculkan jargon-jargon politik. Tentu langkah ini juga akan mendapatkan respon dari masyarakat di media sosial yang sama. Bebasnya ekspresi yang dapat dilakukan di media sosial membuat tak sedikit penggunanya terang-terangan dalam menyampaikan gagasan serta opininya terkait suatu yang terjadi padanya maupun peristiwa-peristiwa yang sedang menjadi tren pada media tersebut. Maka tak mengherankan apabila proses Pilkada DKI menjadi salah satu sasaran empuk bagi pengguna media sosial untuk menyampaikan gagasan serta opini-opininya baik yang bersifat dukungan, celaan hingga provokatif yang kearah SARA.

Salah satu media sosial yang ramai dalam membahas masalah-masalah yang muncul seperti masalah pilkada adalah Twitter. Twitter memiliki 140 karakter ajaib yang dapat di isi dengan *mention* (menandai *user* lain) dan juga *hastag* (topik *tweet*s yang didefinisikan oleh masing-masing *user* dengan memberikan sumbol "#" sebelum topik). *Hastag* inilah yang dapat menjadi kunci dari penggunaan Twitter selama proses pilkada berlangsung. *User* juga memberikan opini mereka dengan menggunakan *hastag-hastag* tertentu yang berhubungan dengan suatu kejadian seperti "#pilkadadki", "#jagajakarta", "#besokgueahok", "#akucoblosaniessandi", dan lain sebagainya.

Mulai tahun 2015 lalu Indonesia telah menjadi salah satu negara yang memiliki pengguna aktif Twitter terbesar didunia (Librianty, 2015). Pada *website*

tersebut Aliza Knox, *Managing Director Online Sales Twitter Asia Pacific*, mengatakan bahwa Indonesia menyumbang rata-rata 500 juta *tweets* setiap hari pada tahun 2014 lalu dengan 3 aktifitas utama yaitu mencari atau membuat konten, menjalin hubungan dengan pengguna lain dan tempat berekspresi. Keterangan yang menunjukkan aktifnya pengguna twitter di Indonesia ini mendukung temuan Ismail (2017) pada saat debat kedua Pasangan Calon Gubernur DKI Jakarta pada jum'at malam tanggal 27 Januari 2017. Ia mengumpulkan semua *tweets* dengan kata kunci "AHY", "AHOK" dan "ANIES" pada pukul 18.00 hingga 22.00 WIB. Hasilnya ialah terdapat sekitar 50.000 *tweets* selama waktu tersebut dengan 55% *tweets* berasal dari luar DKI seperti Bandung, Bogor, Yogyakarta, Bali dan sebagainya. Sisanya (45%) yang berasal dari DKI didominasi oleh pendukung dari kubu Ahok. Hal ini tidak lepas dari banyaknya pengguna twitter populer yang memberikan dukungan terhadap Ahok, sehingga lebih mudah dalam menghimpun *tweets* (Bohang, 2017).

Salah satu kemampuan media digital ialah untuk mendukung formasi ruang publik, dimana keragaman opini dan informasi dapat berinteraksi atau sebaliknya, berfungsi sebagai ruang gema (*echo chamber*) yang memperkuat perspektif dan opini (Colleoni *et. al*., 2014). Ruang gema yang tercipta secara alami di media sosial ini dengan sendirinya akan membentuk kubu-kubu yang saling bertentangan. Sehingga dominasi tren pada media sosial (seperti yang terjadi pada temuan Ismail (2017)) merupakan hasil dari ruang gema yang terjadi secara alami. Hal ini menjadi menarik apabila ternyata ruang gema pada media sosial tersebut mempengaruhi keteguhan masyarakat dalam menentukan pilihannya pada saat pencoblosan.

Serunya proses Pilkada DKI serta perannya media sosial pada proses tersebut membuat peneliti tertarik untuk mengetahui apa saja yang dilakukan pengguna media sosial pada saat pilkada tersebut. Sehingga pada penelitian ini, peneliti menggunakan data *tweets* pada media Twitter terkait dengan Pilkada DKI Putaran 2 untuk menjadi objek penelitian dalam mengaplikasikan metode-metode dalam *text mining*. klasifikasi Naive Bayes dapat memberikan prediksi sentimen yang muncul pada data baru. Namun dalam kasus pilkada, opini yang muncul sangatlah beragam dan

memiliki tren yang berubah-ubah berdasarkan isu yang sedang "dilemparkan" pada masing-masing pasangan calon. Sehingga tak mengherankan apabila prediksi yang dihasilkan sulit untuk mencapai kesesuaian dengan realitanya. Oleh karena itu metode Naive Bayes digunakan untuk menguji sejauh mana batasan sentimen yang dapat diprediksi (oleh metode tersebut). Sedangkan untuk dapat melakukan eksplorasi terhadap *tweets-tweets* tersebut, peneliti menggunakan metode-metode *Unsupervised Learning* seperti *Association Rules*, K-Means dan *Topic Modeling*. Peneliti menggunakan 20.000 data *tweets* yang diambil setiap hari pada tanggal 15-19 April 2017 (masa tenang hingga hari pencoblosan) dengan menggunakan kata kunci "ahok" dan "anies". Pada masa tenang tersebut masyarakat diasumsikan telah memiliki wawasan tentang program-program serta kebijakan yang akan dibentuk oleh masing-masing pasangan calon. Sehingga pada hari-hari tersebut masyarakat mulai meneguhkan pilihan mereka pada saat pencoblosan nanti.

1.2 Rumusan Masalah

Pilkada DKI Jakarta telah menjadi sorotan masyarakat secara nasional. Mengamatinya melalui media sosial menjadi salah satu cara untuk mengetahui isu yang berkembang. Namun kendalanya ialah terlalu banyaknya informasi yang muncul pada media sosial. Sehingga diperlukan suatu metode yang efektif untuk menggambarkan apa yang terjadi. Oleh karena itu peneliti memiliki beberapa rumusan masalah sebagai berikut:

a. Bagaimana efektifitas penggunaan metode *text mining* dalam meringkas informasi apa yang terjadi pada pilkada DKI Putaran 2 pada media sosial Twitter?
b. Apakah metode Naive Bayes mampu melakukan prediksi dengan baik terhadap sentimen pada *tweets* yang muncul?
c. Bagaimana metode *Association Rules* mampu memberikan gambaran yang terjadi pada twitter melalui keterkaitan yang muncul antar kata?

d. Bagaimana hasil pengelompokan yang dilakukan oleh metode K-Means dan *Topic Modeling*?
e. Apakah ada keterkaitan antara pergolakan yang terjadi di media sosial dengan hasil survei lembaga survei dan hasil KPU pada Pilkada DKI Putaran 2?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini ialah menggunakan teknik *text mining* pada data media sosial untuk mengekstrak informasi pada data tweets, melakukan analisis sentimen hingga menemukan keterkaitan antara media sosial Twitter dengan hasil Pilkada DKI Putaran 2.

1.4 Batasan Masalah

Penelitian ini dibatasi pada penggunaan teknik *text mining* pada media sosial Twitter dengan mencari *tweets* menggunakan 2 kata kunci yaitu “ahok” dan “anies”.

BAB 2. TINJAUAN PUSTAKA

2.1 Pilkada DKI Jakarta 2017

Seperti yang dikutip dalam UU No. 8 Tahun 2015 yang menjelaskan bahwa:

> "Pemilihan Gubernur dan Wakil Gubernur, Bupati dan Wakil Bupati, serta Walikota dan Wakil Walikota yang selanjutnya disebut Pemilihan adalah pelaksanaan kedaulatan rakyat di wilayah provinsi dan kabupaten/kota untuk memilih Gubernur dan Wakil Gubernur, Bupati dan Wakil Bupati, serta Walikota dan Wakil Walikota secara langsung dan demokratis"

Pada tahun 2017 ini, pilkada dilakukan serentak di 101 daerah di seluruh Indonesia pada tanggal 15 Februari lalu. Ketua KPU Husni Kamil Manik mengatakan dalam peresmian peluncuran Pilkada Serentak 2017 di Jakarta bulan Februari kemarin bahwa 101 daerah ini terdiri dari 7 provinsi, 76 kabupaten dan 18 kota. Salah satu provinsi yang dimaksud adalah DKI Jakarta (Liputan6, 2017).

Pilkada DKI Jakarta selalu menjadi sorotan media massa, terutama pada saat pemilihan Gubernur dan Wakil Gubernur. Bahkan Direktur Komunikasi Indonesia *Indicator*, Rustika Herlambang menyampaikan kepada *Kompas.com* bahwa terdapat paling tidak ada dua faktor yang menyebabkan Pilkada DKI Jakarta menjadi sorotan dan lebih mendominasi pemberitaan di media *online*. Faktor-faktor tersebut ialah (Fachrudin, 2016):

1. Jakarta merupakan pusat negara. Hal tersebut menjadikan jakarta sebagai acuan bagi daerah lain.
2. Jabatan kepala daerah DKI Jakarta menjadi proyeksi untuk menuju kursi kepresidenan.

Pada 17 Februari lalu, KPU telah mengumumkan hasil resmi rekapitulasi Pilkada DKI Jakarta pada *web* resmi KPU. Hasilnya ialah pasangan Agus-Sylvi memperoleh 17,05% suara, Ahok-Djarot memperoleh 42,91% suara, dan Anies-Sandi memperoleh 40,05% suara (KPU, 2017). Hal ini mengakibatkan Pilkada DKI Jakarta akan melakukan pemilihan putaran kedua. Namun pada pilkada putaran kedua ini

KPU hanya menetapkan dua pasangan calon, yaitu pasangan Ahok-Djarot dan Anies-Sandi. Penetapan kedua pasangan tersebut berdasarkan perolehan dua suara terbanyak pada pilkada putaran pertama lalu. Berdasarkan Keputusan KPU Provinsi DKI Jakarta Nomor : 49/Kpts/KPU-Prov-010/Tahun 2017, KPU memutuskan bahwa jadwal penyelenggaraan Pilkada DKI Jakarta Putaran Kedua mengikuti tahapan berikut :

1. Penyusunan dan penetapan daftar pemilih: 6 Maret - 4 April 2017.
2. Kampanye: 7 Maret-15 April 2017.
3. Kampanye melalui media massa: 9 April-15 April 2017.
4. Masa tenang: 16 April-18 April 2017.
5. Pemungutan suara dan penghitungan suara di TPS: 19 April 2017.
6. Rekapitulasi, penetapan dan pengumuman hasil penghitungan suara tingkat provinsi: 20 April-1 Mei 2017.

2.2 Twitter

Twitter adalah layanan jejaring sosial dan mikroblog *daring* yang didirikan oleh Jack Dorsey pada bulan maret 2006 (Wikipedia, 2016). Memiliki lebih dari 300 juta jiwa pengguna aktif, twitter memberikan layanan kepada penggunanya berupa pesan berisi 140 karakter yang dapat dikirim dan dibaca oleh setiap *followers* yang dikenal dengan sebutan "*tweets*" atau kicauan. Tercatat pada tahun 2007 Twitter memiliki sekitar 5000 kicauan tiap harinya (Kevin, 2010). Namun Twitter tumbuh dengan sangat pesat, pada tahun 2013 Twitter menghasilkan *tweets* hingga mencapai lebih dari 500 juta kicauan setiap harinya (Raffi, 2013).

Twitter API ialah sebuah aplikasi yang diciptakan Twitter agar mempermudah pihak developer lain untuk mengakses informasi *web* Twitter tersebut. API adalah himpunan dari berbagai simbol yang mengekspor dan tersedia untuk pengguna dari suatu *library* untuk dapat melakukan *write* pada aplikasi mereka (Blanchette, 2008). Sebuah API menghubungkan proses, pelayanan, konten dan data suatu developer ke pengguna atau developer lain dengan mudah dan aman (Nijim dan Pagano, 2014).

Dengan menggunakan Twitter API pengguna maupun developer dapat memperoleh berbagai macam informasi baik itu dari salah satu pengguna Twitter ataupun mengetahui tren yang muncul pada saat itu. Sehingga sangat penting penggunaannya untuk memahami dari pengguna-pengguna, *tweets* dan *timeline*.

Twitter API memberikan batasan dalam pengambilan data *tweets* dari suatu pengguna. Addyman (2017) menjelaskan bahwa batas pengambilan maksimum data *tweets* yaitu sekitar 3200 *tweets* dan diambil secara bertahap dengan batas maksimum 200 *tweets* tiap tahapnya.

2.3 Penggalian Data

Analisis data dalam kapasitas yang besar telah menjadi kebutuhan bagi berbagai kalangan baik dari individu, pengusaha, figur publik, hingga partai politik. Salah satu teknik analisis data dalam kapasitas yang besar tersebut adalah *data mining* (penggalian data). Setiap perusahaan telah mendapatkan manfaat dari mengumpulkan dan menggali data seperti, supermarket yang dapat mengetahui produk-produk mana saja yang dibeli secara bersamaan, mesin pencari dapat menentukan urutan alamat *link*, dan rumah sakit dapat mengenal tren serta kelainan pada arsip pasien mereka. Pada dasarnya hasil analisis dari penggalian data tersebut dapat menemukan pola dan keteraturan dalam himpunan data yang sifatnya tersembunyi.

Istilah *data mining* juga dikenal dengan istilah *Knowledge Discovery from Data* atau KDD yang dalam sudut pandang lain *data mining* hanya sebagai langkah penting dalam proses dari penemuan pengetahuan. Proses *data mining* tersebut adalah sebagai berikut (Han *et al.*, 2012):

1. *Cleaning*, proses ini dilakukan untuk menghilangkan data-data yang tidak konsisten dan tidak diperlukan (*noise*).
2. *Integration*, proses ini untuk melihat keterkaitan data dengan data yang lain.
3. *Selection*, pada tahap ini data yang layak untuk dianalisis diambil dari sumber data.

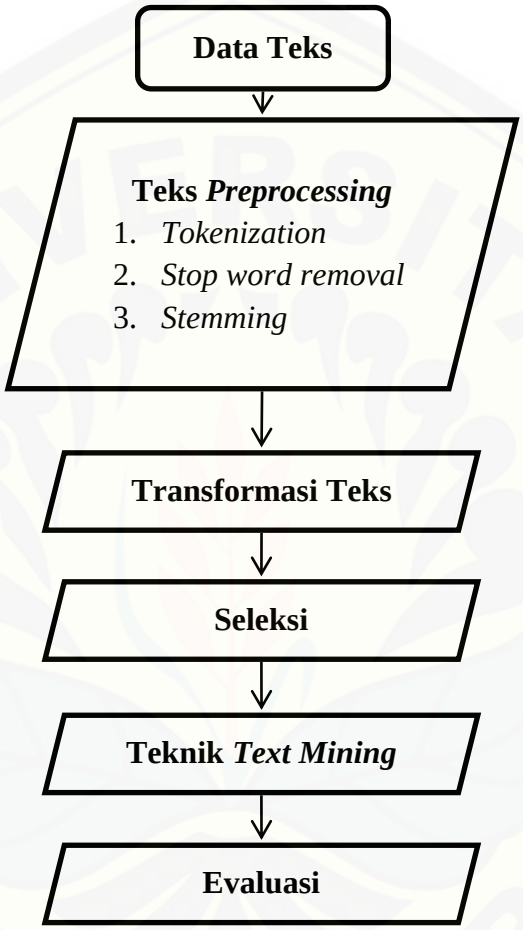
4. *Transformation*, mentransformasi data kedalam format yang sesuai untuk digali kemudian.
5. *Data Mining*, tahap ini adalah proses penting yang digunakan untuk mendapatkan pola data.
6. *Pattern Evaluasi*, setelah ditemukan pola-pola tersebut, selanjutnya ialah memilih pola yang representatif terhadap sumber data.
7. *Knowledge Presentation*, pada tahap ini pola-pola yang representatif divisualisasikan.

Salah satu cabang dari *data mining* adalah *text mining*. *Text mining* merupakan salah satu proses penggalian data yang tidak sama dengan *data mining*. Perbedaan yang paling nampak dari kedua metode ini ialah pada jenis datanya. *Data mining* memproses data-data yang bersifat terstruktur sedangkan *text mining* memproses data-data yang bersifat tidak terstruktur. Data terstruktur ialah data yang dengan mudah untuk diorganisasikan dan umumnya tersimpan dalam *databases*. Sedangkan data yang tidak terstruktur ialah data yang tidak dapat diubah kedalam bentuk data terstruktur, sehingga sulit untuk mengorganisasikannya. Umumnya data ini berupa data teks bahasa alami, gambar, video, email, presentasi, dan lain-lain.

Text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang pengguna berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis (Feldman *et al.*, 2007). Secara umum *text mining* digunakan untuk menunjukan suatu sistem yang menganalisa kuantitas dengan jumlah besar dari teks bahasa alami dan mendeteksi leksikal atau penggunaan pola linguistik dalam usaha untuk mengekstraksi informasi yang berguna (meskipun hanya mungkin benar) (Sebastiani, 2002).

Text mining adalah hal baru dan menarik yang mencoba untuk menyelesaikan masalah informasi yang berlebih dengan menggunakan teknik *data mining, machine learning, natural language processing* (NLP), *information retrieval* (IR), dan *knowledge management* (Feldman *et al.*, 2007). Biasanya *text mining* digunakan untuk melakukan pengkategorian teks, pengelompokan teks, ekstraksi

konsep/kesatuan, analisis sentimen, ringkasan dokumen, dan pemodelan hubungan kesatuan. Terdapat beberapa tahap dalam melakukan analisis *text mining* yang ditunjukan oleh Gambar 2.1 berikut:



Gambar 2.1 Proses *Text Mining*

Berikut adalah penjabaran dari proses Text Mining berdasarkan Gambar 2.1:

1. Data Teks, adalah kumpulan data dokumen yang menjadi objek penelitian seperti yang ditunjukan oleh Gambar 2.2.

```
> Dokumen1
[1] "islam adalah agama perdamaian"
> Dokumen2
[1] "jumlah penduduk Indonesia sangat banyak"
> Dokumen3
[1] "sebagain besar penduduk Indonesia beragama islam"
```

Gambar 2.2 Contoh Data Teks

2. Teks *Preprocessing*, merupakan langkah awal yang berperan untuk menyesuaikan data kedalam format yang diperlukan. Pada langkah ini terdapat 3 tahap yang perlu dilakukan yaitu :
   a. *Tokenization*, tahap ini melakukan pemotongan data kedalam bentuk perkata yang menyusunnya. Seperti yang ditunjukan oleh Gambar 2.3.

```
>
> token1
[1] "islam"      "adalah"     "agama"      "perdamaian"
> token2
[1] "jumlah"    "penduduk"  "indonesia" "sangat"    "banyak"
> token3
[1] "sebagian"  "besar"     "penduduk"  "indonesia" "beragama"  "islam"
>
```

Gambar 2.3 Contoh Proses *Tokenization*

   b. *Stopword removal*, tahap ini melakukan penghapusan pada kata-kata yang tidak diperlukan (kata yang tidak mengandung arti yang representatif). Seperti yang ditunjukan oleh Gambar 2.4

```
> strDok1
[1] "islam"      "agama"      "perdamaian"
> strDok2
[1] "jumlah"    "penduduk"  "indonesia" "banyak"
> strDok3
[1] "sebagian"  "besar"     "penduduk"  "indonesia" "beragama"  "islam"
```

Gambar 2.4 Contoh Proses *Stopword Removal*

   c. *Stemming*, pada tahap ini setiap kata yang memiliki asal kata yang sama disatukan dalam kata yang konsisten. Seperti yang ditunjukan oleh Gambar 2.5.

```
> Dokumen1
[1] "islam agama damai"
> Dokumen2
[1] "jumlah penduduk indonesia banyak"
> Dokumen3
[1] "sebagian besar penduduk indonesia agama islam"
```

Gambar 2.5 Contoh Proses *Stemming*

3. Transformasi teks, langkah ini mentransformasi teks kedalam bentuk dokumen ruang vektor notasi model atau disebut dengan istilah *bag of word* yang berguna untuk memudahkan analisis selanjutnya. Seperti yang ditunjukan oleh Gambar 2.6.

```
          Terms
Docs       agama damai islam banyak indonesia jumlah penduduk besar sebagian
  Dokumen1     1     1     1      0         0      0         0     0        0
  Dokumen2     0     0     0      1         1      1         1     0        0
  Dokumen3     1     0     1      0         1      0         1     1        1
> 
```

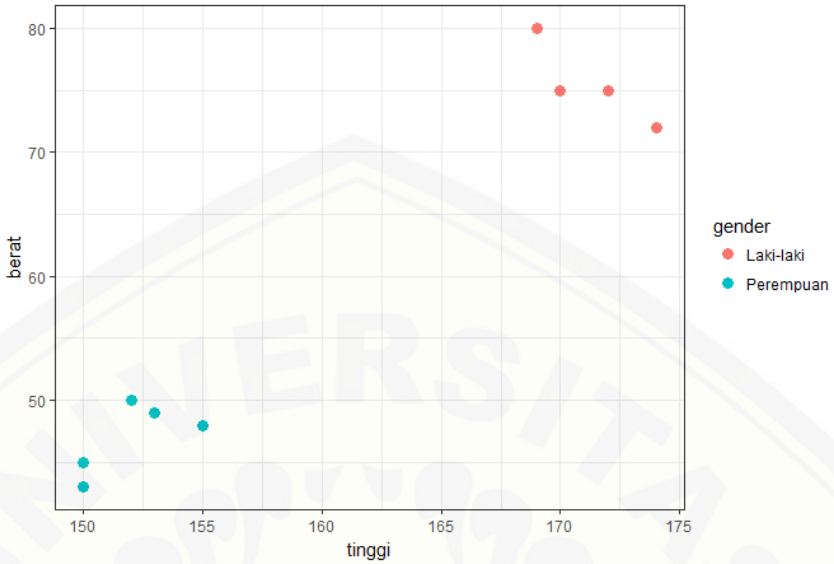
Gambar 2.6 Contoh *bag of word*

4. Seleksi, langkah ini menghapus corak yang dianggap tidak relevan untuk kebutuhan *mining*.
5. Teknik *Text Mining*, terdapat beberapa teknik *text mining* seperti pada *data mining* yaitu pengklasteran, klasifikasi, *information retrieval*, mendeteksi topik, peringkasan, dan ekstraksi topik.
6. Evaluasi, pada langkah ini termasuk dalam evaluasi dan interpretasi hasil.

Namun dalam beberapa literasi, langkah evaluasi tidak dilakukan (Mathiak dan Eckstein, 2004 serta Gaikwad *et al*., 2014). Hal ini disebabkan oleh kebutuhan dari jenis data teks serta pola bahasa yang digunakan. Selain itu, beberapa peneliti menambahkan beberapa tahapan-tahapan pada *text preprocessing* seperti *case folding*, *cleansing* dan *normalization*.

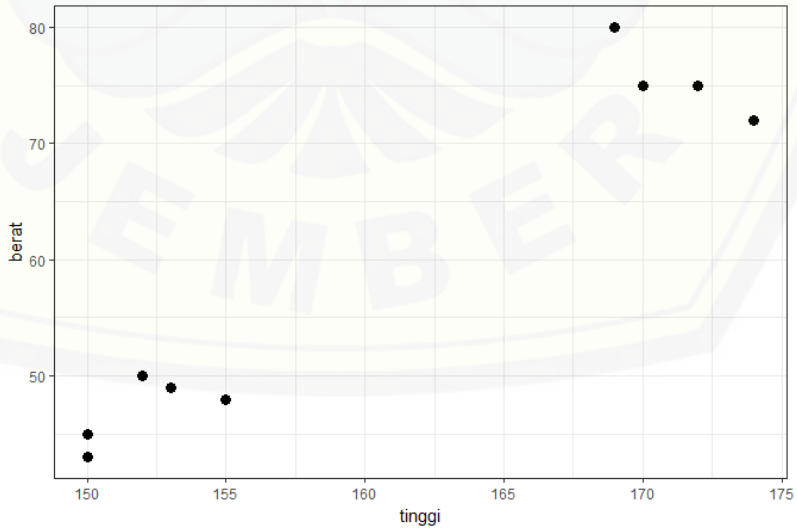
Brownlee (2016) mengatakan bahwa terdapat tiga kelompok data dalam *text mining* yang mana pada setiap data memiliki metode yang berbeda dalam pengerjaannya, ketiga kelompok data tersebut ialah:

1. *Supervised data*, merupakan data teks yang telah memiliki label/pengelompokan. Teknik yang digunakan untuk melakukan analisis pada jenis data ini ialah *Supervised Learning*, dimana pengguna metode ini dapat memprediksi label pada data baru yang tidak terlabeli. Gambar 2.7 berikut adalah contoh visualisasi *supervised data*.



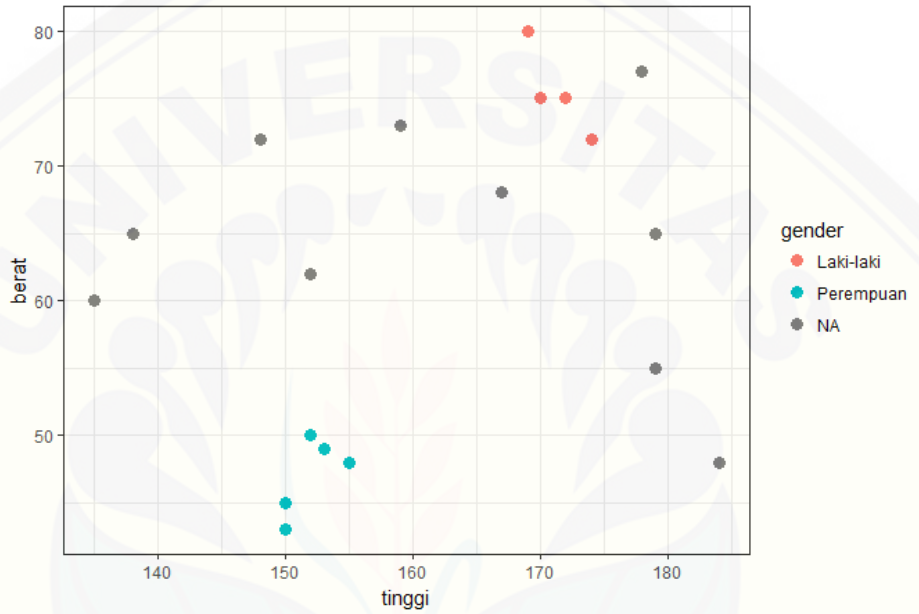
Gambar 2.7 Contoh Plot *Supervised Data*

2. *Unsupervised data*, data ini berkebalikan dengan *supervised data* yaitu data teks yang ada masih belum memiliki label/pengelompokan. Sehingga analisis yang dilakukan ialah sebatas eksplorasi data serta mencari kelompok yang identik berdasarkan karakteristik data ataupun peluang kemunculan kata. Teknik yang digunakan untuk mempelajari data ini ialah *Unsupervised Learning*. Gambar 2.8 berikut adalah contoh visualisasi *unsupervised data*.



Gambar 2.8 Contoh Plot *Unsupervised Data*

3. *Semi-unsupervised data*, data ini merupakan kombinasi dari *supervised data* dan *unsupervised data*. Jadi terdapat beberapa data yang terlabeli dan ada pula yang tidak terlabeli. Dalam melakukan analisisnya tentu memerlukan penggabungan dari teknik *supervised* dan *unsupervised*. Gambar 2.9 berikut adalah contoh visualisasi s*emi-unsupervised data*.



Gambar 2.9 Contoh Plot *Semi-Supervised Data*

2.4 Pembobotan Term

Pembobotan term bertujuan untuk memberikan sebuah nilai pada sebuah term berdasarkan tingkat kepentingan term tersebut didalam sekumpulan dokumen masukan. Pada penelitian ini digunakan metode *Term Frequency – Inverse Term Frequence* (TF-IDF) sebagai proses pembobotan, yaitu dengan cara mencari representasi dari tiap-tiap dokumen dari sekumpulan data *training* dan akan dibentuk menjadi vektor. Salton et al. (2009) merumuskannya sebagai berikut:

$$w(t,d)_{ij} = \frac{tf(t_i, d_j)}{\sum t_{ij}} \times idf \qquad (2.1)$$

$$idf = \log_2\left(\frac{N}{df}\right) \quad (2.2)$$

keterangan:

$tf(t,d)_{ij}$ = kemunculan kata t_i pada dokumen d_j

N = jumlah dokumen pada kumpulan dokumen

df = jumlah dokumen yang mengandung kata t_i

$\sum t_{ij}$ = jumlah kata pada dokumen d_j.

2.5 *Supervised Learning*

Supervised learning merupakan metode yang digunakan untuk memperoleh pembelajaran pada suatu data yang telah memiliki variabel masukan dan variabel luaran (label) untuk melakukan prediksi terhadap data yang hanya memiliki variabel masukan saja. Seperti yang dijelaskan oleh Cunningham et al. (2008) bahwa pembelajaran yang diperoleh dari metode *supervised learning* nantinya akan digunakan untuk melakukan prediksi pada *unseen data*. Disamping itu *supervised learning* adalah metodologi yang paling penting pada mesin pembelajaran (*machine learning*) dan juga memiliki *central importance* dalam memproses data multimedia.

2.5.1 Pelabelan

Dokumen yang terdapat pada internet merupakan data yang tidak terlabeli (*unsupervised data*). Sehingga untuk dapat diproses dengan menggunakan *supervised learning*, diperlukan metode untuk melabeli data-data tersebut. Kendala yang muncul selanjutnya ialah terlalu banyaknya data yang harus diberi label. Oleh karena itu Khushboo et al., (2012) mencoba memberikan label pada dokumen dengan cara menghitung banyak teks yang terkandung pada setiap dokumen. Proses pelabelan dilakukan sebagai berikut (Khushboo *et al.*, 2012):

1. Tentukan kata-kata yang memiliki arti positif juga negatif;
2. Hitung jumlah kata positif dan negatif pada dokumen;

3. Jika jumlah kata positif > jumlah kata negatif, maka label sentimennya adalah positif (skor 1);
4. Jika jumlah kata positif < jumlah kata negatif, maka label sentimennya adalah negatif (skor -1);
5. Jika jumlah kata positif = jumlah kata negatif, maka label sentimennya adalah netral (skor 0).

2.5.2 Klasifikasi Naïve Bayes

Supervised learning umumnya melakukan klasifikasi data terhadap kelas-kelas yang telah ada. Salah satu yang paling sering digunakan untuk melakukan klasifikasi ialah metode Klasifikasi Naïve Bayes. Seperti namanya, proses pengklasifikasian dilakukan berdasarkan peluang yang terdapat pada masing-masing elemen di setiap kelas yang ada. Klasifikasi Naïve Bayes mengasumsikan bahwa kelas untuk klasifikasi adalah independen. Secara sederhana, asumsi tersebut menganggap bahwa adanya fitur khusus pada sebuah kelas yang tidak berhubungan dengan keberadaan fitur lainnya. Contohnya, sebuah buah bila memiliki ciri-ciri berbentuk bulat, berdiameter lebih dari 10 cm dan berwarna hijau akan dianggap sebagai semangka. Walaupun ciri-ciri ini bergantung satu dengan lainnya ataupun atas keberadaan ciri yang lain, semua ciri-ciri tersebut secara independen meningkatkan peluang bahwa buah tersebut adalah semangka.

Metode Klasifikasi Naïve Bayes merupakan salah satu *supervised learning* untuk memberikan klasifikasi terhadap dokumen teks (Li dan Jain, 1998). Bayangkan bahwa suatu dokumen digambarkan dari angka kelas dokumen yang mana dapat dimodelkan sebagai himpunan dari kata-kata dimana peluang (independen) kata yang ke-*i* dari suatu dokumen muncul pada dokumen dari kelas C dapat ditulis sebagai :

$$p(w_i|C)$$

Maka probabilitas bahwa diberikan dokumen D mengandung semua kata w_i, dari kelas C adalah

$$p(D|C) = \prod_i p(w_i|C)$$

Dalam pengklasifikasian teks kita menandai (*tokenize*) dokumen untuk dapat melakukan klasifikasi *in its appropriate class*. Menggunakan aturan pengambilan keputusan dengan "*Max a Posterior Probability*" maka diperoleh pengklasifikasi sebagai berikut (Li dan Jain, 1998):

$$c_{MAP} = \arg\max_{c \in C}\big(P(c|d)\big) = \arg\max_{c \in C}\left(P(c)\prod_{1 \le i \le n} P(w_i|c)\right) \qquad (2.3)$$

keterangan:

d = dokumen

w_i = kata-kata ke-*i* dalam dokumen

c = himpunan dari penggunaan kelas klasifikasi

$P(c|d)$ = *conditional probability* dari kelas c yang diberikan dokumen

$P(c)$ = probabilitas prior dari kelas c

$P(w_i|c)$ = *conditional probability* dari kataw_i diberikan kelas c.

Dapat dilihat bahwa untuk dapat menyimpulkan dokumen tersebut masuk dalam kelas tertentu, harus melakukan estimasi *the product of the probability* dari setiap kata pada dokumen pada kelas khusus (*likelihood*) setelah itu kita kalikan lagi dengan peluang pada kelas khusus (*prior*). Setelah menghitung persamaan (2.3) untuk semua kelas dari himpunan C, selanjutnya dipilih salah satu yang peluangnya paling besar.

2.6 *Unsupervised Learning*

Unsupervised learning merupakan mesin pembelajaran yang khusus digunakan pada data yang hanya memiliki variabel *output* saja. Mesin pembelajaran ini digunakan untuk menemukan pola tersembunyi pada data yang harapannya dapat digunakan untuk menggambarkan karakteristik data. Tidak ada kesimpulan benar ataupun salah pada hasil (*output*) dari *unsupervised learning*. Oleh karena itu, peneliti

harus memahami kajian teori pada topik tertentu yang menjadi objek *unsupervised learning*. Sehingga dengan begitu, peneliti dapat memberikan gambaran yang ideal terhadap *output* dari metode ini.

Metode-metode *unsupervised learning* umumnya melakukan klasifikasi terhadap data. Proses pengklasifikasian dilakukan dengan mempelajari pola data yang saling identik kemudian dilakukan pengelompokan. Beberapa metode yang paling sering digunakan diantaranya ialah *Hierarchical Clustering*, *k-Means Clustering*, *Gaussian Mixture Models*, dan *Topic Modeling*.

2.6.1 *Association Rule*

Association rule adalah salah satu metode yang digunakan untuk menemukan relasi atara term pada data yang sangat besar. *Association rule* dibuat untuk mengidentifikasi relasi yang paling penting pada data. Dalam *association rule* terdapat tiga istilah yaitu, *Support*, *Confidence*, dan *Correlation*. *Support* ialah sebuah indikasi dari seberapa sering term muncul dalam database, sedangkan *Confidence* mengindikasikan berapa kali term yang dimaksud muncul secara kondisional bersamaan dengan term yang lain. Dalam kasus analisis pasar (*market analysis*), *association rule* berguna untuk mempelajari dan memprediksi sikap dari konsumen.

Andaikan diberikan $W = \{w_1, w_2, \dots, w_m\}$ merupakan kumpulan kata. Diberikan database dokumen $D = \{d_1, d_2, \dots, d_n\}$ dimana d_i merupakan dokumen ke-i. Diberikan A dan B adalah kumpulan kata-kata. d_i dikatakan mengandung A jika $A \subseteq d_i$. Maka, *association rule* adalah sebuah implikasi dari formula $A \Rightarrow B$, yang artinya muncul kata A jika dan hanya jika terdapat kata B dimana $A, B \subset W$. (Han *et al.*, 2012).

a. *Support*$(A \Rightarrow B) = P(A \cup B)$.

$$support(A \Rightarrow B) = P(A \cup B) = support(A \cup B) \tag{2.4}$$

b. *Confidence*$(A \Rightarrow B) = P(B|A)$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} \tag{2.5}$$

Secara umum *association rule* memberikan informasi penting dari nilai *support* dan *confidence*. Semakin kecil nilai *support* mengindikasikan semakin kecil pula kemunculan term dalam database. Sedangkan semakin kecil nilai *confidence* mengindikasikan bahwa semakin kecil pula kemunculuan suatu term terhadap term yang lain.

Selanjutnya *assosiation rules* dapat memberikan informasi lanjut untuk dapat melihat korelasi (*correlation*) yang terjadi pada term terhadap term yang lain dengan melihat nilai *Lift* pada kata tersebut dengan kata lainnya. *Lift* dari *association rules* merupakan rasio dari $confidence(A \Rightarrow B)$ terhadap $support(B)$.

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{confidence(A \Rightarrow B)}{support(B)} \quad (2.6)$$

Jika nilai *lift* kurang dari satu, maka muncul kata A berkorelasi negatif terhadap kemunculan kata B. Sedangkan jika *lift* lebih besar dari satu, maka A dan B berkorelasi secara positif yang artinya kemunculan A berimplikasi terhadap kemunculan B (Han *et al.*, 2012).

2.6.2 K-*Means*

Metode klasifikasi k-*Means* merupakan metode yang paling sering digunakan. Hal tersebut disebabkan karena metode ini adalah metode yang cukup sederhana, sehingga dapat dipahami dengan mudah. K-*Means* bertujuan untuk mengelompokan data berdasarkan kemiripan data satu dengan yang lainnya. Berikut adalah algoritma dasar dari k-*Means* (Kumar *et al.*, 2006):

1. Pertama-tama tentukan titik *k* sebagai pusat kelompok yang mungkin terbentuk;
2. Menentukan secara acak posisi *k* yang akan menjadi *centroid*(C_i);
3. Menghitung jarak semua titik (data) ke *k centroids* tersebut lalu memilih jarak terdekat dari masing-masing titik ke *k centroid*. Sehingga setiap titik yang jaraknya paling dekat dengan *centroid* menjadi sebuah kelompok;
4. Menghitung ulang *centroid* dari masing-masing kelompok, hingga *centroid* tidak berubah dengan menggunakan persamaan (2.7).

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x \tag{2.7}$$

keterangan:

c_i = *centroid* baru

m_i = jumlah x yang masuk pada *centroid* C_i.

Prosedur penentuan titik *k* dilakukan secara acak. Setelah itu, menghitung jarak dari masing-masing titik ke *k* titik pusat (*centroid*) dengan menggunakan metode jarak Euclidian yang ditunjukkan oleh persamaan (2.8). Setelah itu proses kembali ketahap 3 hingga nilai *centroid* tidak berubah.

$$d(x_i, x_{i+1}) = \sqrt{\sum_{i=1}^{n_m} (x_i - x_{i+1})^2} \tag{2.8}$$

keterangan:

x_i = nilai pengamatan obyek ke$-i$ dengan $i = 1, 2, \cdots, n_m$

x_{i+1} = nilai pengamatan obyek ke$-(i+1)$

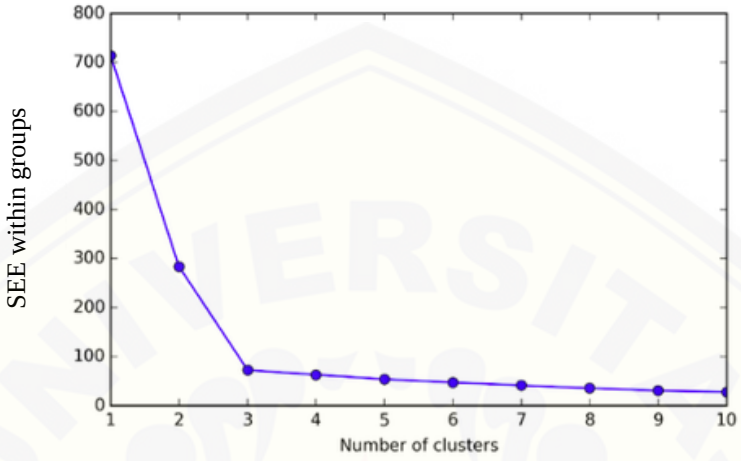
n_m = banyak obyek pada klaster ke$-m$.

Namun dalam menentukan *k* kluster haruslah optimal, karena *k* ditentukan secara manual sehingga sangat memungkinkan apabila *k* kluster yang dipilih terlalu sedikit ataupun terlalu banyak. Salah satu metode yang dapat digunakan untuk membantu menentukan jumlah *k* kluster adalah dengan memvisualisasikan pada plot 2 dimensi atau biasa disebut dengan metode *elbow*. Metode ini bekerja dengan menghitung nilai SSE (*Sum Square of Error*) dari setiap kluster yang terbentuk sebagai berikut (Bholowalia dan Kumar, 2014):

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} dist(c_i, x)^2$$

dengan *dist* adalah fungsi untuk menghitung jarak dengan menggunakan metode jarak Euclid. Setelah itu nilai tersebut digambarkan pada grafik 2 dimensi (*x*, *y*) dimana *x*

adalah jumlah kluster dan *y* adalah SSE masing-masing kluster. Gambar 2.10 berikut adalah contoh grafik dari metode *elbow*:



Gambar 2.10 Contoh Grafik *Elbow* Nilai SSE

Pada Gambar 2.10 diatas maka dapat diputuskan bahwa penggunaan 3 kluster sudah mencapai optimal. Hal tersebut terlihat dari nilai SSE yang mengalami penurunan drastis terjadi sampai pada k-*Means* dengan $k = 3$ dan setelah itu penurun SSE yang terjadi tidak begitu drastis (signifikan).

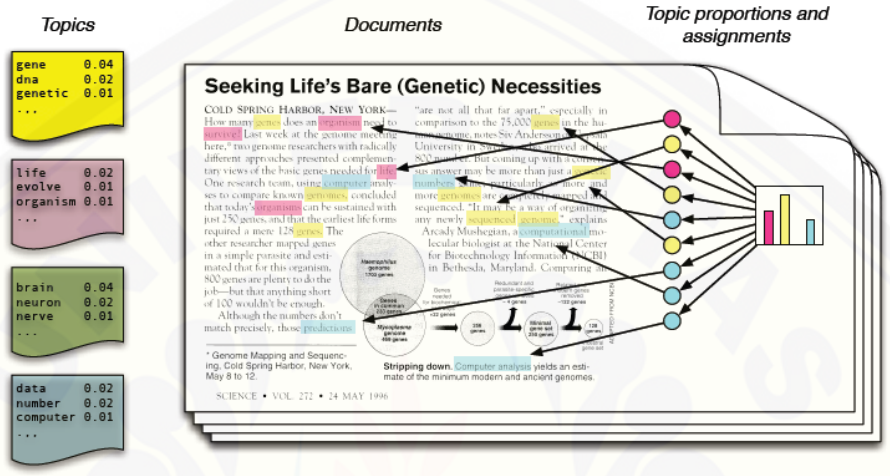
2.6.3 *Topic Modeling* (*Latent Dirichlet Allocation*)

Topic modeling bertujuan untuk menemukan topik dari masing-masing dokumen. Pada dasarnya *topic modeling* menyajikan metode untuk mengorganisasi, memahami dan meringkas kumpulan data teks informasi yang sangat banyak. Kdnuggets.com menjelaskan bahwa *topic modeling* dapat memberikan beberapa manfaat yaitu (Nair, 2016):

1. Menemukan pola topik yang tersembunyi pada kumpulan dokumen.
2. Memberikan anotasi terhadap dokumen sesuai dengan topik.
3. Menggunakan anotasi ini untuk menggorganisasi, mencari dan meringkas teks.

Blei et al. (2003) menyampaikan bahwa *topic modeling* adalah metode statistik yang melakukan analisa kata-kata dari teks asli untuk menemukan tema yang terdapat pada teks tersebut, bagaimana tema tersebut berhubungan satu dengan

lainnya dan bagaimana mereka berubah setiap waktu. Salah satu metode *topic modeling* adalah dengan menggunakan metode *Latent Dirichlet Allocation* (LDA). Intuisi di balik LDA adalah setiap dokumen menunjukan multi topik. Sebagai gambaran Blei et al.(2003) memberikan contoh sebagai berikut:



Gambar 2.11 Contoh teks dokumen (Blei *et al.*, 2003)

Pada Gambar 2.11 diatas, bahwa setiap tanda *highlight* (biru, kuning dan merah muda) merupakan topik-topik yang muncul pada dokumen. Apabila setiap kata pada artikel ditandai demikian, maka akan memiliki sebuah dokumen yang mempunyai proporsi yang berbeda dari masing-masing topik.

LDA mengasumsikan bahwa topik telah terspesifikasi sebelum setiap data terbentuk. Andaikan D adalah kumpulan dokumen *d*, proses pembentukan dokumen berdasarka LDA model adalah sebagai berikut (Darling, 2011):

1. Topik telah terspesifikasi dan telah memiliki distribusi sebagai berikut:

   $$\varphi^{(k)} \sim Dirichlet(\beta), \qquad \text{untuk } k = 1, \dots, K$$

2. Selanjutnya untuk setiap dokumen, kita membentuk kata dengan 2 tahap:
   a. Pilih secara acak sebuah distribusi topik untuk dokumen *d*.

      $$\theta_d \sim Dirichlet(\alpha), \qquad d \in D$$

   b. Setiap kata pada dokumen
      1) Pilih secara acak sebuah topik pada distribusi topik.

$z_i \sim Discrete(\theta_d)$

2) Secara acak pilih sebuah kata dari topik yang berkoresponden pada berdistribusi kosa kata.

$w_i \sim Discrete(\varphi^{(z_i)})$

keterangan:

K = angka dari topik laten pada kumpulan dokumen

$\varphi^{(k)}$ = distribusi probabilitas diskrit pada kosa kata yang merepresentasi distribusi topik ke-*k*

θ_d = distribusi dokumen ke-*d* dari topik yang tersedia

z_i = indeks topik pada kata ke-*i*

w_i = kata ke-*i*

α, β = *hyperparameters* untuk distribusi Dirichlet.

Proses pembuatan dokumen diatas menghasilkan distribusi gabungan sebagai berikut :

$$p(\mathrm{w}, \mathrm{z}, \theta, \varphi | \alpha, \beta) = p(\varphi|\beta)p(\theta|\alpha)p(\mathrm{z}|\theta)p(\mathrm{w}|\varphi_z) \quad (2.9)$$

LDA tidak hanya digunakan untuk melakukan pendeteksian topik saja, namun LDA juga digunakan sebagai salah satu *tools* untuk melakukan analisis *Business Intelligence* pada bank, yaitu untuk mengetahui hubungan antara kebijakan tertentu dengan tren yang dihasilkan (Moro *et al.*, 2014). Selain pada industri perbankan, LDA juga sering digunakan untuk berbagai penelitian lain seperti pada konten percapakapan (Yeh *et al.*, 2014) , bahkan hingga data *software engineering* (Campbell *et al.*, 2014).

Jumlah topik yang optimum pada model LDA dapat ditentukan dengan besar nilai *marginal likelihood* model (2.10) yang dapat diaprosimaksi dengan menggunakan metode *Harmonic Mean* sebagaimana yang ditunjukkan oleh 2.11 berikut:

$$P(\mathrm{w}) = \int P(\mathrm{w}|z)P(z)\, dt \quad (2.10)$$

$$P(\mathrm{w}) = 1/\left[\left(\frac{1}{n}\right)\sum_{i=1}^{T} 1/P(\mathrm{w}|z_i)\right] \quad (2.11)$$

dimana nilai dari $P(w|z_i)$ dapat dihitung dengan menggunakan persamaan 2.12 berikut (Griffiths dan Steyvers, 2004):

$$P(\mathrm{w}|\mathrm{z}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{j=1}^{T} \frac{\prod_{\mathrm{w}}\left(n_j^{(w)} + \beta\right)}{\Gamma\left(n_j^{(.)} + W\beta\right)} \quad (2.12)$$

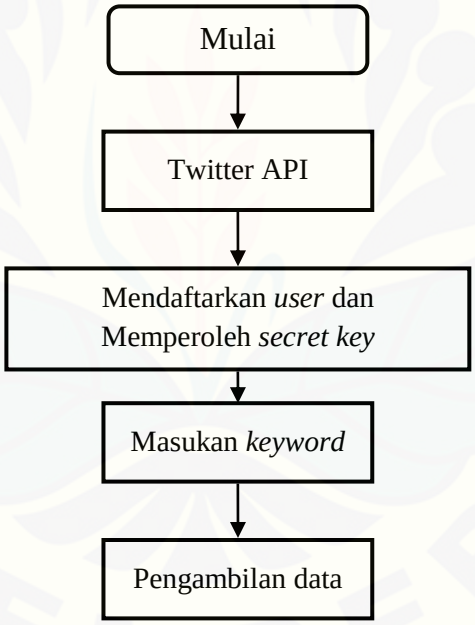
2.7 Visualisasi Data (*Word Cloud*)

Terdapat beberapa cara untuk memudahkan user dalam menyimpulkan hingga menggambarkan karakteristik hingga korelasi data. Diantaranya yang paling sederhana ialah memvisualisasikannya kedalam plot 2 dimensi ataupun 3 dimensi, seperti halnya biplot, boxplot, dan yang biasa digunakan dalam analisis teks atau *text mining* ialah *word cloud*. *Boostlabs.com* menjelaskan bahwa *word cloud* (juga dikenal sebagai *text cloud* atau *tag cloud*) bekerja dengan cara yang sederhana. Kata-kata yang spesifik dimunculkan dengan kondisi semakin banyak teks yang muncul dalam suatu tambang data, maka kata tersebut semakin besar dan tebal. Gambar 2.12 berikut ini adalah salah satu contoh dari *word cloud*.

Gambar 2.12 *Word Cloud* (*Boostlabs.com*)

BAB 3. METODOLOGI PENELITIAN

Penelitian ini berfokus pada eksplorasi data teks dengan menggunakan teknik *text mining*. Data yang digunakan adalah *tweets* yang terkait dengan Pilkada DKI Jakarta Putaran 2. Data *tweets*yang ada tidak dapat diperoleh begitu saja, *user* harus mendaftarkan diri terlebih dahulu ke Twitter API untuk mendapatkan *secret key* (satu *user*, satu *key*). *Secret key* tersebut yang selanjutnya digunakan untuk memperoleh izin dari Twitter agar dapat mengakses/memperoleh *tweets* dari *user* ataupun dari *keyword* tertentu. Proses pengambilan data ditunjukan oleh diagram alir pada Gambar 3.1.



Gambar 3.1 Diagram Alir Proses Pengambilan Data

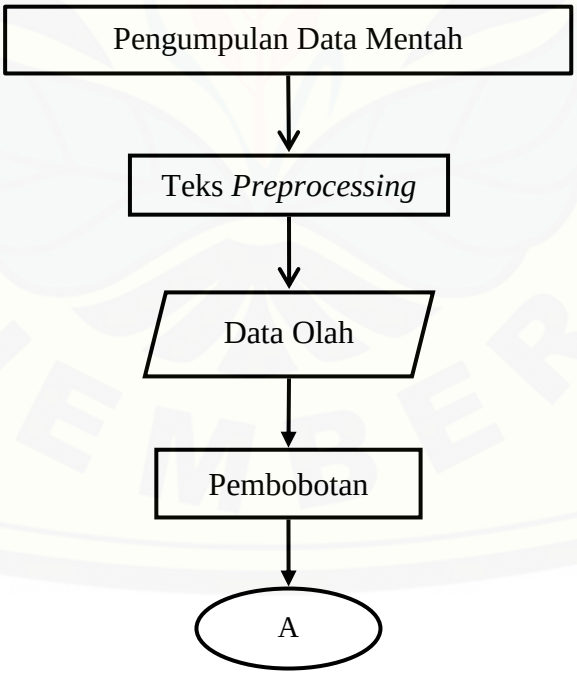
3.1 Data Penelitian

Data penelitian bersumber dari *tweets* pada media sosial Twitter yang berhubungan dengan isu pilkada. Gambar 3.2 berikut ini adalah contoh data yang akan digunakan:

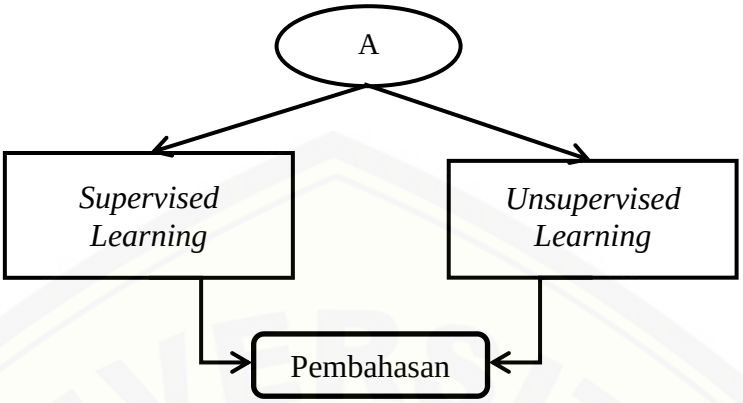
Gambar 3.2 Contoh *Tweets*

3.2 Algoritma

Seluruh data yang telah diambil kemudian disatukan dan dilanjutkan dengan melakukan *Text Preprocessing* (*Case Folding*, *Cleansing*, *Normalization*, *Tokenization*, *Stop word removal*). Setelah itu data diubah kedalam bentuk vektor berdimensi tinggi dan dilanjutkan dengan analisis teks dengan beberapa metode. Diagram alir penelitian ini dapat dilihat pada Gambar 3.3 dan Gambar 3.4 berikut :



Gambar 3.3 Diagram Alir Penelitian



Gambar 3.4 Diagram Alir Penelitian Lanjutan

3.3 Deskripsi Diagram

Data teks pada *tweets* merupakan data tidak terstruktur yang perlu diolah terlebih dahulu agar dapat diproses kedalam teknik *text mining*. Hal tersebut dikarenakan terdapat berbagai macam bahasa yang digunakan serta banyak kata-kata yang memiliki arti sama namun memiliki teks yang berbeda. Misalnya saja kata "Jokowi" dengan "Joko Widodo". Kedua kata tersebut memiliki arti yang sama, sehingga pada kata-kata tersebut akan dipilih salah satu yang dijadikan kata yang mewakili. Untuk itu langkah-langkah seperti yang ditunjukkan oleh diagram alir dilakukan. Berikut ini adalah pemaparan lebih lanjut dari tiap tahapan diagram alir.

3.3.1 Persiapan Data

Pada tahap persiapan data ini peneliti mulai mengambil data dari media sosial Twitter. Proses pengambilan data dilakukan dengan bantuan program. Data yang diperoleh nanti akhirnya akan disimpan kedalam format cvs agar data yang tersimpan dapat digunakan kembali sewaktu-waktu. Berikut adalah proses persiapan data:

1. Pengumpulan Data Mentah

   Data twitter diperoleh dangan memasukan keyword pada search engine Twitter. Terdapat 2 keyword yang digunakan, yaitu "anies" dan "ahok". Setiap data dengan

keyword “anies” dan “ahok” diambil pada tanggal 15-19 April 2017 sebanyak 20.000 data dari pukul 00.00.

2. Teks *Preprocessing*

Tidak semua data yang masuk menjadi kebutuhan dalam pengolahan. Sehingga sebelum proses pengolahan data maka terdapat beberapa langkah awal yang dilakukan agar pada saat proses analisa nanti menjadi lebih efektif dan efisien. Berikut adalah langkah-langkah yang dilakukan:

a. Case Folding

Tahap pertama yang dilakukan ialah *case folding*, dimana pada tahap ini semua huruf kapital pada dokumen teks diubah menjadi huruf kecil. Hal ini bertujuan untuk menghindari dua kata yang sama dianggap berbeda oleh program karena perbedaan huruf kapital saja. Contoh hasil tahapan ini dapat dilihat pada Tabel 3.1 berikut:

Tabel 3.1 Proses *Case Folding*

Data Awal	Data Setelah *Case Folding*
RT @pppjakarta: 22.H. Djarot mendukg setiap kebijakn Ahok mewujudkan visi & misi saat Pemilukada DKI 2012, ykni membangn Jakarta Baru #SiaP...	rt @pppjakarta: 22.h. djarot mendukg setiap kebijakn ahok mewujudkan visi & misi saat pemilukada dki 2012, ykni membangn jakarta baru #siap...
"RT @VIVAcoid: Survei SPIN: Anies-Sandi 52 Persen, Ahok-Djarot 43 Persen https://t.co/QgsWbgGEvt	"rt @vivacoid: survei spin: anies-sandi 52 persen, ahok-djarot 43 persen https://t.co/qgswbggevt
RT @VisiMuslimNews: HTI Dilawan, AHOK Kok Dikawan ? | Read More : https://t.co/YHwcuZq98E https://t.co/U2Mz6Ke1Zt	rt @visimuslimnews: hti dilawan, ahok kok dikawan ? | read more : https://t.co/yhwcuzq98e https://t.co/u2mz6ke1zt

b. Cleansing

Tahapan ini bertujuan untuk membersihkan kata-kata dari tanda baca atau simbol-simbol lainnya yang dikenal dengan istilah *noise*. *Noise* merupakan

suatu bentuk data yang nantinya akan mengganggu proses pengolahan data tersebut. *Noise* tersebut diantaranya ialah alamat link, kata yang diawali dengan karakter '@' (*mention*) dan karakter '#' (*hastag*) pada *tweets*, serta simbol ataupun tanda baca lainnya. Tabel 3.2 berikut adalah contoh hasil tahapan *cleansing*:

Tabel 3.2 Proses *Cleansing*

Data Awal	Data Setelah *Cleansing*
rt @pppjakarta: 22.h. djarot mendukg setiap kebijakn ahok mewujudkan visi & misi saat pemilukada dki 2012, ykni membangn jakarta baru #siap...	djarot mendukg setiap kebijakn ahok mewujudkan visi amp misi saat pemilukada dki ykni membangn jakarta baru
"rt @vivacoid: survei spin: anies-sandi 52 persen, ahok-djarot 43 persen https://t.co/qgswbggevt	survei spin aniessandi persen ahokdjarot persen
rt @visimuslimnews: hti dilawan, ahok kok dikawan ? | read more : https://t.co/yhwcuzq98e https://t.co/u2mz6ke1zt	hti dilawan ahok kok dikawan read more

c. Tokenization

Tahap *tokenization* ini melakukan pemotongan kata dalam dokumen teks menjadi potongan kata-kata tunggal. Tujuan dari tahapan ini agar proses-proses selanjutnya menjadi lebih mudah seperti penghitungan kata, pembobotan hingga transformasi kedalam bentuk vektor berdimensi tinggi. Tabel 3.3 berikut adalah contoh hasil dari proses *tokenization*:

Tabel 3.3 Proses *Tokenization*

Data Awal	Data Setelah *Tokenization*
djarot mendukg setiap kebijakn ahok mewujudkan visi amp misi saat pemilukada dki ykni membangn jakarta baru	'djarot' 'mendukg' 'setiap' 'kebijakn' 'ahok' 'mewujudkan' 'visi' 'amp' 'misi' 'saat' 'pemilukada' 'dki' 'ykni' 'membangn' 'jakarta' 'baru'

Data Awal	Data Setelah *Tokenization*
survei spin aniessandi persen ahokdjarot persen	‘survei’ ‘spin’ ‘anies’ ‘sandi’ ‘persen’ ‘ahok’ ‘djarot’ ‘persen’
hti dilawan ahok kok dikawan read more	‘hti’ ‘dilawan’ ‘ahok’ ‘kok’ ‘dikawan’ ‘read’ ‘more’

d. Normalization

Keterbatasan karakter yang diberikan twitter mengakibatkan terdapat beberapa kata yang sengaja disingkat oleh pengguna agar *tweet* mencangkup opini penggunanya. Selain itu juga terdapat beberapa kata yang memiliki arti sama, sehingga kata-kata dengan arti sama tersebut diseragamkan oleh peneliti. Begitu pula dengan kata-kata dengan bahasa asing akan diubah kedalam bahasa Indonesia. Oleh karena itu proses *normalization* ini dilakukan dengan tujuan untuk dapat meminimalisir pengulangan-pengulangan kata yang memiliki arti sama. Tabel 3.4 berikut ini menunjukan beberapa kata singkatan serta kata yang memiliki arti yang sama. Sedangkan pada Tabel 3.5 menunjukan contoh hasil normalisasi data:

Tabel 3.4 Contoh Daftar Beberapa Kata-kata yang di Normalisasi

Kata Singkatan		Kata yang Memiliki Arti Yang Sama	
Kata	Ganti	Kata	Ganti
badja	ahok	byk	banyak
basuki	ahok	*New*	baru
busuki	ahok	Gini	begini
ahokbersa	ahokbersamasyiah	gitu	begitu
ahokbersam	ahokbersamasyiah	bener	benar
ahoksyiahbersama	ahokbersamasyiah	aberagam	beragam
ahoker	ahokers	bera	beragam
ahokerlah	ahokers	berdo	berdoa
		stop	berhenti
		bkn	bukan
		nyela	cela
		dlm	dalam
		dpt	dapat

Tabel 3.5 Proses *Normalization*

Data Awal	Data Setelah *Normalization*
‘djarot’ ‘mendukg’ ‘setiap’ ‘kebijakn’ ‘ahok’ ‘mewujudkan’ ‘visi’ ‘amp’ ‘misi’ ‘saat’ ‘pemilukada’ ‘dki’ ‘ykni’ ‘membangn’ ‘jakarta’ ‘baru’	‘djarot’ ‘mendukung‘setiap’ ‘kebijakan’ ‘ahok’ ‘mewujudkan’ ‘visi’ ‘amp’ ‘misi’ ‘saat’ ‘pemilukada’ ‘jakarta’ ‘yakni’ ‘membangun’ ‘jakarta’ ‘baru’
‘survei’ ‘spin’ ‘anies’ ‘sandi’ ‘persen’ ‘ahok’ ‘djarot’ ‘persen’	‘survei’ ‘spin’ ‘anies’ ‘sandi’ ‘persen’ ‘ahok’ ‘djarot’ ‘persen’
‘hti’ ‘dilawan’ ‘ahok’ ‘kok’ ‘dikawan’ ‘read’ ‘more’	‘hti’ ‘dilawan’ ‘ahok’ ‘kok’ ‘dikawan’ ‘baca’ ‘lanjut’

e. Stopword Removal

Setelah proses *normalization*, selanjutnya ialah menghapus kata-kata (*stopwords removal*) yang tidak diperlukan dalam penelitian. Tabel 3.6 berikut adalah daftar *stopwords* yang digunakan merujuk dari Tala (2005). Sedangkan Tabel 3.7 adalah hasil dari proses *stopwords removal* tersebut :

Tabel 3.6 Contoh Daftar *Stopwords* (Tala, 2015)

No.	Kata	No.	Kata	No.	Kata
1	ada	11	antaranya	21	bagaimanapun
2	adalah	12	apa	22	bagi
3	adanya	13	apaan	23	bagian
4	adapun	14	apabila	24	bahkan
5	agak	15	apakah	25	bahwa
6	agaknya	16	apalagi	26	bahwasanya
7	agar	17	apakah	27	baik
8	akan	18	artinya	28	bakal
9	akankah	19	asal	29	bakalan
10	akhir	20	asalkan	30	balik

Tabel 3.7 Proses *Stopwords Removal*

Data Awal	Data Setelah *Stopwords Removal*
‘djarot’ ‘mendukung’ ‘setiap’ ‘kebijakan’ ‘ahok’ ‘mewujudkan’	‘ djarot’ ‘ mendukung’ ‘kebijakan’ ‘ahok’ ‘mewujudkan’ ‘visi’ ‘misi’

Data Awal	Data Setelah *Stopwords Removal*
‘visi’ ‘amp’ ‘misi’ ‘ saat’ ‘pemilukada’ ‘ jakarta’ ‘ yakni’ ‘membangun’ ‘ jakarta’ ‘ baru’	‘pemilukada’ ‘jakarta’ ‘membangun’ ‘jakarta’ ‘baru’
‘survei’ ‘spin’ ‘anies’ ‘sandi’ ‘persen’ ‘ahok’ ‘djarot’ ‘persen’	‘survei’ ‘spin’ ‘anies’ ‘sandi’ ‘persen’ ‘ahok’ ‘djarot’ ‘persen’
‘hti’ ‘dilawan’ ‘ahok’ ‘kok’ ‘dikawan’ ‘baca’ ‘lanjut’	‘hti’ ‘dilawan’ ‘ahok’ ‘kok’ ‘dikawan’ ‘baca’ ‘lanjut’

3. Data Olah

Setelah semua tahap tersebut dilakukan, maka diperoleh data baru yang siap untuk di olah dengan metode-metode *text mining*. Tabel 3.8 berikut adalah sampel data olah setelah melalui tahap-tahap *preprocessing* diatas:

Tabel 3.8 Hasil *Preprocessing*

Data Mentah	Data Olah
RT @pppjakarta: 22.H. Djarot mendukg setiap kebijakn Ahok mewujudkan visi & misi saat Pemilukada DKI 2012, ykni membangn Jakarta Baru #SiaP...	djarot dukung bijak ahok wujud visi misi pemilukada jakarta bangun jakarta baru siap
"RT @VIVAcoid: Survei SPIN: Anies-Sandi 52 Persen, Ahok-Djarot 43 Persen https://t.co/QgsWbgGEvt	survei spin anies sandi persen ahok djarot persen
RT @VisiMuslimNews: HTI Dilawan, AHOK Kok Dikawan ? | Read More : https://t.co/YHwcuZq98E https://t.co/U2Mz6Ke1Zt	hti lawan ahok kok kawan bacalanjut

3.3.2 Pembobotan *Term Frequency – Inverse Term Frequence* (TF-IDF)

Pembobotan *term* merupakan matriks yang representatif terhadap kumpulan dokumen yang dapat digunakan untuk berbagai hal seperti pengklasifikasian serta pemilihan *term* yang memiliki nilai kuat (bobot yang besar) pada kumpulan

dokumen. Matriks ini direpresentasikan sebagai kumpulan fitur dan dapat diilustrasikan sebagai $d_j = [w_{ij}, w_{2j}, \dots, w_{kj}]$ dengan d_j merupakan dokumen ke-j dan w_{kj} merupakan kata ke-k pada dokumen ke-j. Matriks tersebut yang umum disebut dengan *Term Document Matrix* (Matriks Dokumen *Term*) berisi nilai-nilai kemunculan fitur, dengan baris sebagai dokumen ke-j dan kolom sebagai kata ke-k. Sebagai gambaran, berikut ini adalah contoh data tweet dalam matriks TF-IDF

Tweet 1 : partai indonesia bhineka persatuan pilkada

Tweet 2 : ahok bhineka persatuan pilkada putaran dua

Tweet 3 : ahok persatuan putaran anies ungguli

Tweet 4 : partai pdip ppp ahok

Tweet 5 : survei anies ungguli ahok

Kemudian data tweet tersebut dibentuk kedalam matriks *term* sebagaimana yang ditunjukan oleh Tabel 3.9 berikut :

Tabel 3.9 *Document Term Matrix*

Kata	Dokumen				
	1	2	3	4	5
bhineka	1	1	0	0	0
indonesia	1	0	0	0	0
partai	1	0	0	1	0
persatuan	1	1	1	0	0
pilkada	1	1	0	0	0
ahok	0	1	1	1	1
dua	0	1	0	0	0
putaran	0	1	1	0	0
anies	0	0	1	0	1
ungguli	0	0	1	0	1
pdip	0	0	0	1	0
ppp	0	0	0	1	0
survei	0	0	0	0	1

Selanjutnya dengan menggunakan persamaan (2.1) dan (2.2) diperoleh matriks bobot TF-IDF sebagai mana yang ditunjukan oleh Tabel 3.10 berikut :

Tabel 3.10 Matriks TF-IDF

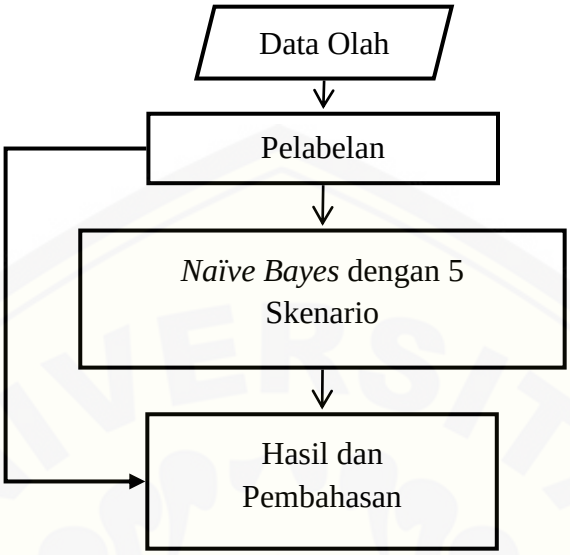
Kata	Dokumen				
	1	2	3	4	5
bhineka	0,2644	0,2203	0	0	0
indonesia	0,4644	0	0	0	0
partai	0,2644	0	0	0,3305	0
persatuan	0,1474	0,1228	0,1474	0	0
pilkada	0,2644	0,2203	0	0	0
ahok	0	0,0537	0,0644	0,0805	0,0805
dua	0	0,3870	0	0	0
putaran	0	0,2203	0,2644	0	0
anies	0	0	0,2644	0	0,3305
ungguli	0	0	0,2644	0	0,3305
pdip	0	0	0	0,5805	0
Ppp	0	0	0	0,5805	0
survei	0	0	0	0	0,5805

3.3.3 Analisis

Setelah data diperoleh dan disimpan, maka akan mudah untuk dipanggil kembali sesuai dengan kebutuhan analisis. Proses analisis data terbagi menjadi 3 yaitu menemukan *pattern* serta memvisualisasikannya, menentukan serta membaca sentimen yang muncul pada setiap dokumen tersebut, dan melakukan pengkelompokan data berdasarkan karakteristik data ataupun topiknya. Berikut adalah proses analisis yang dilakukan:

1. *Supervised Learning*

   Data olah yang telah diperoleh kemudian akan dilakukan pelabelan berdasarkan jumlah kata positif dan negatif. Setelah itu data yang telah terlabeli akan digunakan untuk menguji metode Klasifikasi Naïve Bayes dengan menggunakan 5 skenario. Tahap-tahap pengerjaan digambarkan oleh diagram alir Gambar 3.5 berikut:



Gambar 3.5 Diagram Alir *Supervised Learning*

2. *Unsupervised Learning*

   Teknik *unsupervised learning* dapat mengelompokkan data berdasarkan karakteristiknya hingga mengelompokan berdasarkan topik yang tersembunyi. Setelah teknik ini dilakukan peneliti hanya mampu melakukan eksplorasi terhadap kelompok-kelompok maupun topik-topik yang terbentuk. Selain itu dengan metode *Association rules* kita dapat melihat korelasi antar kata yang saling berkaitan. Secara keseluruhan penelitian ini digambarkan oleh diagram alir pada Gambar 3.6 berikut:

BAB 5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini berfokus terhadap penggunaan beberapa metode-metode dalam *text mining* untuk mememaksimalkan penemuan informasi yang terdapat dalam data yang sangat banyak. Dalam aplikasinya, metode-metode *text mining* dapat digunakan dalam melakukan analisis sentimen hingga menemukan topik-topik yang terdapat pada data. Selain itu dengan melihat keterkaitan yang terjadi antar kata dapat memberikan petunjuk lain yang memberikan informasi tambahan terkait dengan data. Berikut adalah kesimpulan-kesimpulan yang didapat dalam menggunakan metode-metode *text mining* pada data *tweets* dalam khasus Pilkada DKI Putaran 2 kemarin :

1. Visualisasi data dengan menggunakan metode *wordcloud* dapat memberikan pokok pebahasan yang muncul pada data yang sangat banyak. Keunggulan *wordcloud* dalam memunculkan kata berdasarkan jumlah kata yang paling sering muncul cukup efektif dalam meringkas data pada kasus Pilkada DKI. Selain itu dengan meneliti proses *preprocessing* dapat menghilangkan kata-kata yang tidak bermakna, sehingga kesimpulan yang diambil tidak terganggu dengan kemunculan kata-kata yang tidak bermakna tersebuts. Namun metode ini menjadi tidak efektif apabila tidak ada kata yang mendominasi pada data *tweets*.
2. Metode Naive Bayes dalam melakukan prediksi sentimen pada data *tweets* “anies” dan “ahok” di hari yang sama memiliki tingkat keakuratan rata-rata sebesar 80%. Namun apabila hasil pembelajaran yang dilakukan hari ini digunakan untuk memprediksi hari esok ataupun hari lain, metode Naive Bayes mengalami penurunan keakuratan yang cukup besar yaitu hanya mampu memprediksi dengan keakuratan rata-rata 50% kebawah.
3. *Association Rules* dapat memberikan informasi yang lebih jelas dengan melihat besar nilai *Lift* serta *Confidence* pada suatu kata dengan kata-kata yang lain. Setiap kata yang memiliki besar *Lift* yang lebih dari 1 terhadap kata lain mengindikasikan bahwa kedua kata tersebut saling mendukung kemunculannya. Dengan informasi tersebut, peneliti dapat lebih mudah menyimpulkan apa yang sedang menjadi pembahasan pada data tersebut.
4. Pengelompokan data berdasarkan metode *k*-Means cenderung memerlukan jumlah *centroid* yang besar. Namun besarnya jumlah centroid tersebut masih memiliki pola yang sama yaitu

ada 1 kelompok yang memiliki anggota yang dominan daripada kelompok-kelompok yang lain.

5. *Topic Modeling* berhasil memberikan pengelompokan yang seragam pada topik yang terbentuk. Setiap kelompok yang dibentuk oleh *Topic Modeling* memiliki keanggotaan yang cukup merata. Selain itu, sebagaian besar data *tweets* berhasil dikelompokan secara optimum oleh *Topic Modeling* dengan menggunakan 75 topik. Dengan begitu peneliti lebih mudah dalam mengeksplorasi topik yang muncul dibandingkan dengan pengelompokan *k*-Means yang cenderung memilik 1 kelompok yang keanggotaannya mendominasi.
6. Terdapat dugaan-dugaan bahwa tren yang terjadi pada media sosial Twitter memiliki keterkaitan atau bahkan mempengaruhi hasil pilkada. Hal ini ditunjukkan dari tren sentimen yang muncul pada Twitter sejalan dengan hasil pilkada, baik itu pada sentimen yang positif maupun sentimen yang negatif.

5.2 Saran

Pengolahan data yang sangat banyak semakin lama semakin sering digunakan, baik pada bidang sains, ekonomi hingga sosial. Penelitian ini hanya mampu memberikan gambaran bagaimana cara kerja pengolahan data yang sangat banyak tersebut pada khasus data teks yang masih memerlukan perbaikan-perbaikan serta tambahan-tambahan untuk dapat mendekati sempurna. Oleh karenanya, untuk penelitian selanjutnya peneliti menyarankan beberapa hal berikut:

1. Eksplorasi metode-metode pembelajaran Supervised dan Unsupervised yang lain untuk dapat menemukan kekuatan serta kelemahan metode-metode tersebut. Selain itu dengan memberikan visualisasi yang beragam yang sesuai dengan kondisi data, dapat mempermudah dalam membaca tren data tersebut.
2. Untuk menguji metode pemberian sentimen yang telah dilakukan, bisa dengan menggunakan data yang telah memiliki sentimen kemudian diberikan sentimen kembali untuk dilihat hasilnya.
3. Selain pengolahan data teks yang tidak terstruktur, pengolahan data terstruktur juga menarik untuk dipelajari. Terutama jika ingin mengembangkan kemampuan analisis pada bidang-bidang ekonomi dan bisnis.

4. Untuk dapat menemukan keterkaitan yang lebih jauh lagi, sebaiknya memperbesar rentang pengambilan data. Karena dengan menggunakan data pada 5 hari peneliti masih belum dapat memberikan kesimpulan yang pasti.

DAFTAR PUSTAKA

Addyman, C. 2017. *Easily Download All Tweets From A User*. http://www.craigaddyman.com/mining-all-tweets-with-python/ [diakses pada 6 Maret 2017]

Bholowalia, P., dan A. Kumar. 2014. EBK-Means: *A Clustering Technique based on Elbow Method and K-Means in WSN*

Blanchette, J. 2008. *The Little Manual of API Design*. *Trolltech, a Nokia company*

Blei, D. M. Andrew, Y. Ng., Michael, I. J. 2003. *Latent Dirichlet Allocation*. *Journal of Machine Learning Research*.

Bohang, F., K. 2017. Temuan Mengejutkan Keramaian Twitter di Debat Kedua Cagub DKI. http://tekno.kompas.com/read/2017/01/28/18143657/temuan.mengejutkan.keramaian.twitter.di.debat.kedua.cagub.dki. [9 Juni 2017]

Boostlabs.com.*Word Clouds & the Value of Simple Visualizations*. http://www.boostlabs.com/what-are-word-clouds-value-simple-visualizations/. [diakses 21 Maret 2017]

Brownlee, J. 2016. *Supervised and Unsupervised Machine Learning Algorithms*. http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. [diakses 6 Maret 2017]

Campbell, J. C., Hindle,A., Stroulia,E. 2014. *Latent Dirichlet Allocation*: *Extracting Topics from Engineering Data*.

Colleoni, E., Rozza, A., *Arvidsson A. 2014. Echo Camber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data*. *Journal of Communication 64*. 317-322

Cunningham, P., Cord, M., Delany, S. J. 2008. *Supervised Learning*. *Journal of Cognitive Technologies* pp 21-49.

Darling, W. M. 2011. *A Theoretical and Pratical Implementation Tutorial on Topic Modeling and Gibbs Sampling*.

Fachrudin, F. 2016. Ini Alasan Pemberitaan Pilkada DKI Jakarta Mendominasi. http://nasional.kompas.com/read/2016/12/02/23151391/ini.alasan.pemberitaan.pilkada.dki.jakarta.mendominasi. [diakses 20 Maret 2017]

Feldman, R., Sanger, James. 2007. *The Text Mining Handbook. Cambrige Universiy Press.*

Gaikwad, S. V., Chaugule, A., Patil, P. 2014. *Text Mining Methods and Techniques*. *International Journal of Computer Applications*, vol. 85, no. 17, pp. 42-45

Han, J., Kamber, M., and Pie. 2012. *Data Mining : Concepts and Techniques*. 3rded. USA : Elsevier Inc.

Kevin, W. 2010. *Measuring Tweets. Twitter Official Blog*. [diakses 6 Maret 2017]

Khushboo, N., Swati , T., Vekariya, K., Shailendra, M. 2012. *Mining of Sentences Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm. int. J. Computer Technology & Applications*, Volume 3, pp. 987991

KPU. 2017. Hasil Hitung TPS (Form C1) Provinsi Dki Jakarta. https://pilkada2017.kpu.go.id/hasil/t1/dki_jakarta. [diakses 20 Maret 2017]

Kpu.go.id. 2017. https://pilkada2017.kpu.go.id/hasil/2/t1/dki_jakarta [9 Juni 2017]

Kumar, V., Steinbach, M., Tan, P-N. 2006. *Introduction to Data Mining*.

Li, Y. H., Jain, A. K. 1998. *Classification of Text Document*. *The Computer Journal*, vol. 41, no. 8.

Librianty**,** A. 2015. Ini yang Dilakukan Orang Indonesia di Twitter. http://tekno.liputan6.com/read/2217704/ini-yang-dilakukan-orang-indonesia-di-twitter. [9 Juni 2017]

Liputan6. 2017. Ini 101 Daerah yang Gelar Pilkada Serentak 2017. http://pilkada.liputan6.com/read/2436435/ini-101-daerah-yang-gelar-pilkada-serentak-2017. [diakses 20 Maret 2017]

Mathiak, B., Eckstein, S. 2004. *Five Steps to Text Mining in Biomedical Literature*. *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics*. 24 September 2004

Moro, S., Cortez, P., Rita, P.2014. *Business intelligence in banking*: *A literature analysis from* 2002 to 2013. Elsevier.

Nailufar, N., N. 2017. Ini Hasil Rekapitulasi Suara Putaran Kedua Pilkada DKI Jakarta. http://megapolitan.kompas.com/read/2017/04/30/06030941/ini.hasil.rekapitulasi.suara.putaran.kedua.pilkada.dki.jakarta [9 Juni 2017]

Nair, G. 2016. *Text Mining 101*: *Topic Modeling*. http://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html. [diakses 28 Maret 2017]

Nijim, S., Pagano, B. 2014. *API for Dummies*. *USA* : John Wiley & Sons, Inc.

Raffi, K. 2013. *New Tweets per second record, and how!*.*Twitter Official Blog*.[diakses 6 Maret 2017]

Ridwan. 2017. Hasil Survei Pilkada DKI: 7 Lembaga Survei Menangkan Anies, 1 Unggulkan Ahok. http://pojoksatu.id/pilkada-dki-jakarta-2017/2017/04/15/hasil-survei-pilkada-dki-7-lembaga-survei-menangkan-anies-1-unggulkan-ahok/ [9 Juni 2017]

Salton, G., Buckley, C. 1988. *Term-weighting approaches in automatic text retrieval*. *Information Processing and Management*, **24**/5, 513--523.

Sebastiani, F. 2002. *Machine learning in automated text categorization*.*ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47

Steyvers, M., Griffiths, T. L. 2004. *Finding Scientific Topics*. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS).

Tala, F., Z. 2015. *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*. M.S. thesis. M.Sc. Thesis. Master of Logic Project. Institute for Logic, Language and Computation. Universiteti van Amsterdam The Netherlands.

Wallach, H., M, Murray , I, Mimno, D. 2009. *Evaluation methods for topic models*. *In Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada.

Wikipedia. 2016. Twitter. https://id.wikipedia.org/wiki/Twitter. [diakses 6 Maret 2017]

Yeh, J.-F., Tan, Y.-S., Lee, C.-H., Yu, L.-C. 2014. *Topic Model Allocation of Conversational Dialogue Records by Latent Dirichlet Allocation*.