



ANALISIS SENTIMEN PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES (STUDI KASUS PEMILIHAN GUBERNUR DKI JAKARTA 2017)



© Hak cipta milik IPB (Institut Pertanian Bogor)

MUHAMMAD HADIYAN RASYADI



**DEPARTEMEN ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2017**

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi berjudul Analisis Sentimen pada Twitter Menggunakan Metode Naïve Bayes (Studi Kasus Pemilihan Gubernur DKI Jakarta 2017) adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

 Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2017

Muhammad Hadiyan Rasyadi
NIM G64130027

Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



ABSTRAK

MUHAMMAD HADIYAN RASYADI. Analisis Sentimen pada Twitter Menggunakan Metode Naïve Bayes (Studi Kasus Pemilihan Gubernur DKI Jakarta 2017). Dibimbing oleh HUSNUL KHOTIMAH.

Pemilihan umum Gubernur DKI Jakarta 2017 menjadi salah satu topik yang ramai diperbincangkan di Twitter. *Tweet* dan *retweet* yang beredar pada saat pemilu mengandung opini dari masyarakat kepada setiap pasangan calon Gubernur dan Wakil Gubernur DKI Jakarta. Penelitian ini memprediksi sentimen masyarakat kepada pasangan calon Gubernur dan Wakil Gubernur DKI Jakarta putaran kedua. Data yang diperoleh adalah *tweet* yang menyebut kata kunci @AhokDjarot dan @JktMajuBersama. Pengambilan data menggunakan *library* *Tweepy* dengan bahasa pemrograman Python 2.7 dilakukan pada tanggal 4, 12, 16, 17, 18, dan 19 April 2017. Penelitian ini membagi sentimen menjadi 3 kelas, yaitu positif, negatif dan netral. Pelabelan dilakukan secara manual kemudian dilakukan pemodelan menggunakan metode Naïve Bayes pada *library* *NLTK*. Pembuatan model menggunakan 400 data latih dan pengujian model menggunakan 200 data uji. Hasil penelitian diperoleh akurasi data uji sebesar 60.60% dengan tingkat sensitifitas tertinggi terdapat pada kelas positif dan spesifisitas tertinggi terdapat pada kelas netral.

Kata Kunci: klasifikasi, Naïve Bayes, Pemilu DKI Jakarta 2017, Twitter

ABSTRACT

MUHAMMAD HADIYAN RASYADI. Twitter Sentiment Analysis Using the Naïve Bayes Method (Case Study of Governor Election in DKI Jakarta 2017). Supervised by HUSNUL KHOTIMAH.

The 2017 Jakarta General Election is one of the most frequently discussed topics on Twitter. Wide spread tweet and retweet on general election contains opinions from the public to every pair of candidates for governor and vice governor of DKI Jakarta. This study predicts the sentiments of society to the candidate in the second round. The obtained data are tweets that mention @AhokDjarot and @JktMajuBersama. The data collection using a *Tweepy* library with Python 2.7 programming language was done on 4, 12, 16, 17, 18, and 19 April 2017. This research divides sentiment into 3 classes, which are positive, negative and neutral. The labeling was done manually and the modeling was using Naïve Bayes method in *NLTK* library. For modeling 400 training data are used and for testing 200 test data are used. The result of this research are the accuracy of test data is 60.60% with the highest sensitivity level is positive class and the highest specificity is neutral class.

Keywords: classification, Jakarta election 2017, naïve bayes, Twitter

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



ANALISIS SENTIMEN PADA TWITTER MENGGUNAKAN METODE NAÏVE BAYES (STUDI KASUS PEMILIHAN GUBERNUR DKI JAKARTA 2017)

©

Hak cipta milik IPB (Institut Pertanian Bogor)

MUHAMMAD HADIYAN RASYADI

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer
pada
Departemen Ilmu Komputer

**DEPARTEMEN ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2017**

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



© Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural U

Pengujit:

- 1 Dean Apriana Ramadhan, SKomp MKom
- 2 Dr. Yani Nurhadryani, SSi MT

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Judul Skripsi: Analisis Sentimen pada Twitter Menggunakan Metode Naïve Bayes (Studi Kasus Pemilihan Gubernur DKI Jakarta 2017)
Nama : Muhammad Hadiyan Rasyadi
NIM : G64130027



Hak cipta milik IPB (Institut Pertanian Bogor)

Bogor Agricultural University

Disetujui oleh

Husnul Khotimah, SKomp MKom
Pembimbing

Diketahui oleh



Dr Ir Agus Buono, MSi MKom
Ketua Departemen

Tanggal Lulus: 31 AUG 2017

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *subhanahu wa-ta'ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan dan atas izin dan kehendak-Nya penulis dapat ditunjukkan kemudahan-Nya. Topik yang dipilih dalam penelitian yang dilaksanakan sejak bulan November 2016 ini ialah analisis data Twitter, dengan judul Analisis Sentimen pada Twitter Menggunakan Metode Naïve Bayes (Studi Kasus Pemilihan Gubernur DKI Jakarta 2017).

Penyusunan dan penyelesaian tugas akhir ini tidak terlepas dari bantuan berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada:

- 1 Bapak dan Ibu tercinta dan tersayang, yaitu Bapak Suranto SE dan Ibu Dra Sri Herlina. Kakak tercinta dan tersayang Nurul Hidayah SSi. Mereka lah orang orang hebat yang selalu mengalirkan do'a, memberikan semangat dan motivasi yang tinggi.
- 2 Ibu Husnul Khotimah, SKomp MKom selaku pembimbing.
- 3 Dean Apriana Ramadhan, SKomp MKom dan Dr. Yani Nurhadryani, SSi MT selaku penguji sidang.
- 4 Royan, Faldhi, Fajar, Ajmal, Ivan, Hadi, dan Haekal terimakasih untuk semangat dan motivasi yang kalian berikan selama di PT Kontrakan.
- 5 Ajeng Rafii, Akhiyar, Elfakar, Ryan, Miftah, Punto, Mutia, dan Aulia serta teman teman Sonic 5.0 yang selalu menjadi perantara jalan keluar dari Allah *subhanahu wa ta'ala*.
- 6 Shintia Hawari yang telah membantu penyusunan skripsi.
- 7 Kak Fauzan, Kak fahmi, Kak Kiki, Winda, Adi, Caca, dan tema teman Creative Media Serum G 1437 terimakasih atas pengalaman yang kalian ajarkan.
- 8 Lorong Tiga Asrama TPB C3 yang selalu merusuhkan suasana.
- 9 Seluruh dosen, staf tata usaha, dan staf pegawai Departemen Ilmu Komputer IPB. Semoga dukungan, bantuan, bimbingan dan motivasi yang diberikan menjadi amal jariah dan diberikan ganjaran yang berlebih oleh Allah *subhanahu wa ta'ala*.

Semoga karya ilmiah ini bermanfaat.

Bogor, Agustus 2017

Muhammad Hadiyan Rasyadi

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



DAFTAR ISI

DAFTAR TABEL	viii
DAFTAR GAMBAR	viii
DAFTAR LAMPIRAN	viii
PENDAHULUAN	1
Latar Belakang	1
Perumusan Masalah	2
Tujuan Penelitian	2
Manfaat Penelitian	2
Ruang Lingkup Penelitian	2
METODE	3
Data Penelitian	3
Tahapan Penelitian	3
Lingkungan Pengembangan	7
HASIL DAN PEMBAHASAN	7
Pengumpulan dan Pelabelan Data	7
Pembagian Data	9
Praproses Data	9
Pemodelan Klasifikasi	11
Pengujian	12
Evaluasi Hasil	13
Prediksi Sentimen	14
SIMPULAN DAN SARAN	17
Simpulan	17
Saran	17
DAFTAR PUSTAKA	17
RIWAYAT HIDUP	19

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



DAFTAR TABEL

1 Confusion matrix table	6
2 Data Twitter yang terkumpul	8
3 Contoh record data berupa Json	8
4 Hasil pelabelan data secara manual	8
5 Hasil pembersihan data	10
6 Contoh hasil normalisasi	10
7 Contoh hasil stemming data dan penghapusan pungtuasi	11
8 Contoh hasil penghapusan stopword	11
9 Hasil pengujian model	12
10 Hasil confusion matrix table	13
11 Hasil sensitifitas dan spesifisitas setiap kelas	13

DAFTAR GAMBAR

1 Tahapan proses sentimen analisis masyarakat terhadap calon cagub dan cawagub berdasarkan tweet dan retweet pada Pemilu tahun 2017	3
2 Tahap praproses data	4
3 Wordcloud dari hasil praproses setiap kelas (a) Kelas positif (b) Kelas negatif (c) Kelas netral	11
4 Model classifier yang paling berpengaruh untuk menentukan kelas	12
5 Hasil Total Prediksi Sentimen	14
6 Hasil sentimen pasangan calon nomor 3 @JktMajuBersama	14
7 Hasil Sentimen Pasangan Nomor 2 @AhokDjarot	15
8 Perbandingan kelas positif dari 2 akun pasangan calon	15
9 Perbandingan kelas negatif dari 2 akun pasangan calon	16
10 Perbandingan kelas netral dari 2 akun pasang calon	16

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



PENDAHULUAN

Latar Belakang

Media sosial adalah sebuah layanan yang memfasilitasi dalam pertukaran informasi dan topik secara berkelanjutan dengan cakupan yang luas (Schrape 2016). Salah satu jenis media sosial yang populer di kalangan pengguna Internet adalah *microblogging*. Pengguna Internet berpindah dari *blog* atau *mailing list* menuju ke *microblogging* karena akses dan format penulisan pesannya yang mudah (Pak dan Paroubek 2010). Contoh layanan *microblogging* adalah Twitter. Berdasarkan data dari *website statistica*¹, perkembangan pengguna Twitter pada tahun 2011 mencapai 117 juta pengguna dan pada tahun 2017 mencapai 328 juta pengguna aktif di seluruh dunia. Pengguna Twiiter di Indonesia pada tahun 2016 menurut *website socialmemos*² mencapai 29 juta pengguna dengan 2.4% dari 10 juta *tweet worldwide*. Pengguna Internet menuliskan opini dan pendapat tentang berbagai topik pada layanan Twitter. Karena pengguna mengekspresikan tentang berbagai topik seperti politik, hal ini menjadikan Twitter sebagai sumber data yang berpotensi dan efisien mengenai Pemilihan Gubernur DKI Jakarta. Penelitian ini menggunakan Twitter karena memiliki beberapa keuntungan, yaitu digunakan oleh berbagai kalangan pengguna, memiliki pesan singkat disebut *tweet* yang mengandung opini masyarakat yang beragam, bertambah setiap saat, dan persebaran berita yang cepat.

Twitter mulai digunakan untuk kepentingan politik oleh masyarakat atau institusi politik (Stieglitz dan Xuan 2012). Twitter berperan aktif dalam proses komunikasi antara institusi politik dengan masyarakat terutama pada saat kampanye berlangsung. Kampanye merupakan salah satu rangkaian Pemilu. Komunikasi antara calon Gubernur dan Wakil Gubernur kepada masyarakat berlangsung secara intensif (Mahendra dan Oka 2014). Selama masa kampanye, partai politik atau calon gubernur dan calon wakil gubernur memerlukan usaha publisitas yang tinggi, salah satu caranya melalui Internet (Situmorang 2013). Menurut Undang-undang Peraturan Komisi Pemilihan Umum RI (PKPU) Nomor 6 Tahun 2016 mengatur bahwa DKI Jakarta dalam menentukan Gubernur dan Wakil Gubernur melalui Pemilihan Umum Kepala Daerah (Pilkada) dan mendapatkan minimal 50% suara untuk memenangkan pada putaran pertama. Putaran kedua diadakan ketika tidak ada calon yang mendapatkan minimal 50% suara dan diikuti oleh calon yang memperoleh suara terbanyak pertama dan kedua. Pemilu DKI Jakarta 2017 dilakukan 2 putaran, karena pada putaran pertama tidak ada calon yang mendapatkan 50% suara.

Pada penelitian sebelumnya yang dilakukan oleh Wang *et al.* (2012), peneliti tersebut membangun sebuah sistem untuk melakukan analisis secara *real-time* pada pemilihan presiden Amerika Serikat tahun 2012. Pembangunan sistem menganalisis tentang keterhubungan antara pemilu dan pengaruhnya terhadap opini yang ada di masyarakat. Sistem tersebut menampilkan total data yang telah terkumpul, sentimen, dan kata yang sering muncul pada 5 menit terakhir dari semua akun kandidat. Proses pengumpulan data menggunakan sistem yang dibangun oleh

¹ <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

² <http://socialmemos.com/social-media-statistics-for-indonesia/>



InfoSphere Streams platform dari IBM dan pembuatan model menggunakan Amazon Mechanical Turk dengan metode Naïve Bayes. Mengacu pada sistem tersebut, penelitian ini bertujuan mengetahui sentimen analisis Twitter calon Gubernur (cagub) dan calon Wakil Gubernur (cawagub) DKI Jakarta tahun 2017. Proses pengumpulan data yang dilakukan pada putaran kedua menggunakan Twitter API pada *library* Tweepy. Pengumpulan data dilakukan berdasarkan kata kunci berupa akun calon Gubernur dan Wakil Gubernur DKI Jakarta tahun 2017, dengan jumlah tertentu dan pengambilan sampel di tanggal tertentu. Selama pengumpulan data, peneliti mengambil atribut isi teks, tanggal dan *id user* kemudian menambahkan atribut sentimen dan akun. Setelah itu tahap praproses, yaitu normalisasi data dengan cara mengubah kata tidak baku menjadi baku, menghilangkan angka, menghilangkan tanda baca dan simbol, *stemming data* dengan menghilangkan imbuhan pada setiap kata, dan penghapusan *stopword*. Kemudian masuk ke tahap prediksi sentimen menggunakan metode Naïve Bayes, untuk dihitung jumlah sentimen negatif, sentimen positif, dan sentimen netral.

Perumusan Masalah

Rumusan permasalahan pada penelitian ini adalah dibutuhkannya analisis tentang opini publik pada pasangan cagub dan cawagub Pilkada DKI Jakarta 2017. Proses analisis dilakukan berdasarkan *tweet* yang melakukan *mention* kepada akun resmi cagub dan cawagub.

Tujuan Penelitian

Tujuan dari penelitian ini adalah melakukan analisis sentimen terhadap calon Gubernur dan Wakil Gubernur DKI Jakarta tahun 2017 menggunakan metode Naïve Bayes.

Manfaat Penelitian

Manfaat dari penelitian ini yaitu mengetahui sentimen analisis masyarakat terhadap pasangan cagub (calon gubernur) dan cawagub (calon wakil gubernur) DKI Jakarta 2017. Sentimen dibagi menjadi 3 kelas, yaitu kelas positif, kelas negatif, dan kelas netral. Selain itu, hasil analisis yang didapat digunakan untuk mengetahui pengaruh media sosial Twitter pada Pilkada DKI Jakarta.

Ruang Lingkup Penelitian

Lingkup dari penelitian ini, yaitu:

- 1 Akun yang digunakan adalah akun yang resmi terdaftar di KPU (Komisi Pemilihan Umum).
- 2 Pengambilan data dilakukan pada Pilkada DKI Jakarta 2017 di putaran kedua.

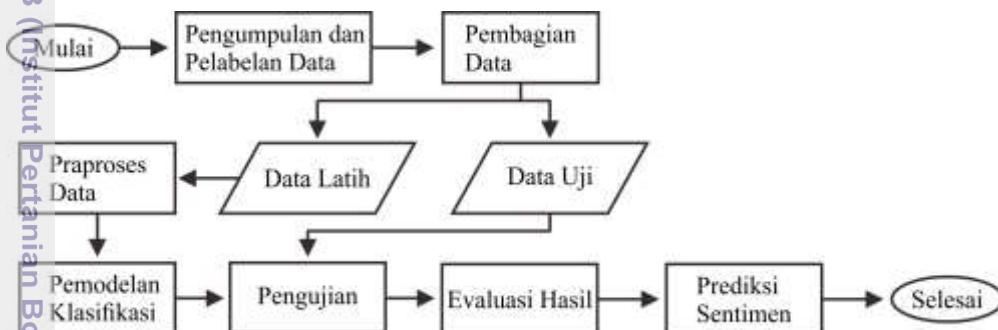
METODE

Data Penelitian

Data yang digunakan adalah *tweet* yang melakukan *mention* kepada akun cagub dan cawagub Pemilu DKI Jakarta tahun 2017. Pengambilan sampel data *tweet* dilakukan secara acak pada tanggal tertentu. Pengambilan sampel data menggunakan kata kunci dan dibatasi sebanyak 4 000 *tweet* per hari. Kata kunci yang digunakan berdasarkan akun Twitter resmi, meliputi @AhokDjarot untuk pasangan nomor 2 dan @JktMajuBersama untuk pasangan nomor 3.

Tahapan Penelitian

Proses sentimen analisis pada pasangan cagub dan cawagub berdasarkan *tweet* dan *retweet* dapat dilihat pada Gambar 1. Tahap pertama, peneliti melakukan pengumpulan dan pelabelan data, kemudian tahap praproses, tahap pemodelan klasifikasi, tahap pengujian, dan tahap evaluasi hasil. Kemudian peneliti melakukan prediksi sentimen terhadap data yang telah dikumpulkan.



Gambar 1 Tahapan proses sentimen analisis masyarakat terhadap calon cagub dan cawagub berdasarkan *tweet* dan *retweet* pada Pemilu tahun

Pengumpulan dan Pelabelan Data

Proses pengambilan data menggunakan *library* Tweepy. Penggunaan *library* ini membutuhkan akses berupa OAuth 1 yang bisa didapat dari *website* Twitter Developer. Penggunaan OAuth 1 pada *library* ini mendapatkan izin untuk melakukan akses menggunakan API yang tersedia. *Library* ini memanfaatkan API Twiiter Berupa *search* yang bisa mencari *tweet* yang memiliki kecocokan dengan kata kunci yang diberikan. Peneliti memberikan kata kunci berupa akun Twitter @AhokDjarot dan @JktMajuBersama. Jadwal Pilkada DKI Jakarta 2017 putaran kedua pada masa kampanye dan debat publik adalah tanggal 17 Maret hingga 15 April 2017, masa tenang adalah tanggal 16 hingga 18 April 2017, dan pemungutan suara pada tanggal 19 April 2017. Pengumpulan data *tweet* dilakukan dengan cara mengambil sampel pada hari Selasa tanggal 4 April 2017, hari Rabu tanggal 12 April 2017, hari Minggu, Senin, dan Selasa tanggal 16, 17, dan 18 April 2017, serta hari Rabu tanggal 19 April 2017. Persebaran penarikan sampel data dilakukan agar mewakili rangkaian Pemilu DKI Jakarta 2017. Data yang diambil berupa 3 atribut

- Hak Cipta Dilindungi Undang-Undang
© Hak cipta milik IPB (Institut Pertanian Bogor)
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



- yaitu *id_user*, *created_at* dan *text*. Kemudian peneliti menambah atribut sentimen dan akun untuk mempermudah dalam penelitian. Data *tweet* yang terkumpul disimpan dalam format JSON.

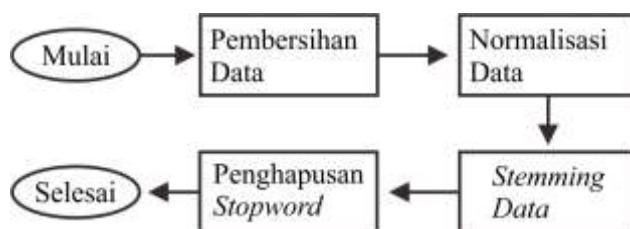
Proses pelabelan data dilakukan dengan cara mengambil 400 data *tweet* yang teratas di tanggal 4 April 2017 pada akun @AhokDjarot dan 200 data *tweet* dari akun @JktMajuBersama dengan melakukan penghapusan data yang redundant. Kemudian peneliti memberikan label pada setiap *tweet* secara manual. Label yang diberikan ada 3 kelas, yaitu kelas positif, kelas negatif, dan kelas netral. Kelas positif merupakan kelas yang berisikan data yang mengandung kata bermakna positif, pernyataan setuju, dan dukungan. Kelas negatif merupakan kelas dengan data yang mengandung kata bermakna negatif, ejekan, dan kontra. Kelas netral merupakan kelas yang mengandung berita.

Pembagian Data

Data yang telah melalui proses pelabelan, dilakukan pembagian data menjadi 2, yaitu data latih dan data uji. Data latih berjumlah 400 data dan berfungsi untuk membangun model awal. Data uji berjumlah 200 data dan digunakan untuk melakukan pengujian terhadap model yang telah terbentuk.

Praproses Data

Data latih yang terkumpul masih berbentuk data yang belum terstruktur dengan isi dari setiap *tweet* masih dalam bahasa yang tidak baku. Tahap ini bertujuan untuk menghilangkan karakter yang tidak relevan dan mengurangi kualitas model selain itu juga dapat meningkatkan kualitas data latih. Gambar 2 menjelaskan tentang tahap praproses data yang dilakukan pada penelitian.



Gambar 2 Tahap praproses data

Praproses data pada penelitian ini mengalami 4 tahapan yaitu:

1 Pembersihan Data

Proses ini menghilangkan karakter yang mengurangi kualitas data latih, seperti *link*, nama akun Twitter seseorang dan tanda *hashtag* (#). Program akan mencari karakter yang telah tentukan, kemudian karakter tersebut akan dihapus dari data tweet tersebut.

2 Normalisasi Data

Tahap ini berfungsi untuk mengubah kata yang tidak baku menjadi kata baku dan menghilangkan karakter yang berulang (Aziz 2013). Proses normalisasi data memiliki kumpulan kata yang tidak baku beserta pasangan kata baku. Proses normalisasi dilakukan dengan mencari kata yang tak baku atau kata

- Hak Cipta Dilindungi Undang-Undang**

 1. Dilarang mengutip sebagian atau seluruh karya tulis jika:
 - a. Pengutipan hanya untuk kepentingan pendidikan
 - b. Pengutipan tidak merugikan kepentingan yang dimiliki oleh pengarang
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis tanpa izin pengarang.

ini tanpa mencantumkan dan menyebutkan sumber: n, penelitian, penulisan karya ilmiah, penyusunan laporan : 152

ran, penulisan kritik atau tinjauan suatu masalah.

singkatan pada teks menggunakan fungsi *regex*, kemudian kata tersebut digantikan oleh pasangan kata yang baku dan tidak disingkat.

3 *Stemming Data*

Proses ini berupa penghilangan imbuhan dan akhiran pada setiap token. Proses ini menggunakan *library* PySastrawi berdasarkan algoritme Nazief dan Adriani dengan menghilangkan imbuhan dari sebuah kata menjadi kata dasar (Asian 2007). Algoritme ini meliputi aturan bahasa yang kompleks, terdapat imbuhan, ambiguitas, *overstemming*, *understemming*, bentuk jamak, serapan, akronim dan yang lainnya. Sastrawi memiliki 40 aturan pemenggalan yang digunakan untuk melakukan proses *stemming*. Kamus kata dasar yang digunakan oleh PySastrawi berasal dari website <http://kateglo.com/> pada 23 September 2009 dengan ada sedikit perubahan yang mereka lakukan. Selain itu, pada *library* ini juga telah tersedia penghapusan pungtuasi. Sehingga proses ini selain melakukan *stemming data*, juga sebagai penghapusan pungtuasi. Penghapusan pungtuasi terjadi bersamaan dengan proses *stemming*, ketika menemukan pungtuasi akan dihapuskan.

4 Penghapusan *Stopword*

Tahapan ini menghilangkan kata yang tidak deskriptif atau tidak penting. Kata ini perlu dihilangkan karena tidak mengandung atau merepresentasikan data. *Stopword* merupakan kata-kata dengan frekuensi kemunculan yang tinggi (Tala 2003). *Stoplist* yang digunakan berasal dari analisis kata dasar yang dilakukan oleh Tala (2003).

Pemodelan Klasifikasi

Pemodelan klasifikasi diawali dengan memisahkan setiap kata dari data yang telah selesai melalui tahap praproses. Kemudian dilakukan perhitungan frekuensi kemunculan setiap kata pada seluruh dokumen. Seluruh kata tersebut akan dicek kemunculannya pada setiap data. Berdasarkan informasi tersebut, dihitung peluang munculnya suatu kata pada kelas tertentu. Hasil dari perhitungan peluang tersebut dijadikan sebagai model dalam proses memprediksi menggunakan *library* NLTK. Penelitian ini menggunakan metode Naïve Bayes sebagai analisis sentimen yang digunakan. Algoritme pada *library* ini menggunakan aturan Bayesian untuk mengekspresikan $P(\text{label} | \text{fitur})$ dalam hal $P(\text{label})$ dan $P(\text{fitur} | \text{label})$. Bayesian Klasifikasi merupakan klasifikasi secara statistik dengan memprediksi suatu data terprediksi ke dalam kelas tertentu (Han *et al.* 2012). Berikut adalah aturan Bayesian yang digunakan pada *library* NLTK.

$$P(\text{label} | \text{fitur}) = \frac{P(\text{label}) * P(\text{fitur} | \text{label})}{P(\text{fitur})} \quad (1)$$

Library ini membuat asumsi bahwa semua fitur bersifat independen, sehingga:

$$P(\text{label} | \text{fitur}) = \frac{P(\text{label}) * P(f_1 | \text{label}) * P(f_2 | \text{label}) * P(f_3 | \text{label}) * \dots * P(f_n | \text{label})}{P(\text{fitur})} \quad (2)$$

- Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
- Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Algoritme ini hanya menghitung pembilang untuk setiap label, dan menormalkannya menjadi satu daripada menghitung peluang dari fitur, seperti berikut:

$$P(\text{label}|\text{fitur}) = \frac{P(\text{label}) * P(f_1|\text{label}) * P(f_2|\text{label}) * P(f_3|\text{label}) * \dots * P(f_n|\text{label})}{\text{SUM}[\text{label}] (P(f_1|\text{label}) * P(f_2|\text{label}) * P(f_3|\text{label}) * \dots * P(f_n|\text{label}))} \quad (3)$$

Pengujian

Pengujian model menggunakan data uji yang berjumlah 200 data *tweet*. Model akan memprediksi setiap *tweet* ke dalam suatu kelas. Penentuan kelas ini berdasarkan model yang terbentuk dari data latih.

Evaluasi Hasil

Proses evaluasi pada penelitian ini berdasarkan 3 ukuran, yaitu akurasi, sensitifitas dan spesifisitas yang dihitung berdasarkan persamaan 4, 5, dan 6. Nilai pada persamaan tersebut dihitung berdasarkan pada *confusion matrix* pada Tabel 1.

Tabel 1 *Confusion matrix table*

		Prediksi	
		Positif	Negatif
Aktual	Positif	<i>True Positive</i> (TP)	<i>False Negative</i> (FN)
	Negatif	<i>False Positive</i> (FP)	<i>True Negative</i> (TN)

Proses evaluasi dimulai dari pencarian akurasi. Akurasi merupakan persentase dari suatu kelas terprediksi dengan benar oleh model yang sudah dibuat (Han *et al.* 2012). Rumus akurasi ditunjukkan pada persamaan 4, dengan P adalah semua data aktual di kelas positif dan N adalah semua data aktual di kelas negatif.

$$\text{Akurasi} = \frac{TP + TN}{P + N} \quad (4)$$

Pada penelitian ini, proses evaluasi yang dilakukan selain akurasi adalah menghitung spesifisitas dan sensitifitas setiap kelas. Menurut Han *et al.* (2012), sensitifitas merupakan tingkat seberapa baik program memprediksi data ke dalam kelas yang sesuai dengan kelas aktualnya, sedangkan spesifisitas bisa disebut sebagai persentase sebuah program memprediksi sebuah data ke bukan kelas aktualnya. Sensitifitas dan spesifisitas dapat dilihat pada persamaan berikut ini:

$$\text{Sensitifitas} = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{Spesifisitas} = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (6)$$

Prediksi Sentimen

Proses ini melakukan prediksi terhadap data yang telah dikumpulkan. Data yang digunakan adalah data yang belum mengalami proses pelabelan. Hasil prediksi dibandingkan untuk setiap paslon dalam tiga kelas, yaitu kelas positif,

kelas negatif, dan kelas netral kemudian ditampilkan kedalam sebuah diagram garis dan dilakukan perbandingan antara data dari akun @AhokDjarot dengan data dari akun @JktMajuBersama.

Lingkungan Pengembangan

Spesifikasi perangkat keras dan perangkat lunak yang digunakan untuk penelitian ini adalah sebagai berikut:

- 1 Perangkat keras menggunakan laptop dengan spesifikasi:

- Processor Intel i7-4750HQ.
- RAM 4GB.
- 1TB HDD.

- 2 Perangkat lunak yang digunakan yaitu:

Python 2.7 sebagai bahasa pemrograman pada penelitian.

Library Tweepy untuk mengambil data *tweet* menggunakan API Twitter.

Library NLTK untuk memprediksi sentimen.

Flask untuk pembuatan API pada tampilan web.

Postman untuk melakukan cek pada API.

HASIL DAN PEMBAHASAN

Pengumpulan dan Pelabelan Data

Penelitian ini menggunakan Python 2.7 sebagai bahasa pemrograman dan library Tweepy untuk melakukan pengumpulan data. Tweepy dapat mengumpulkan data menggunakan Twitter API dengan OAuth 1 sebagai akses. Peneliti mendapatkan OAuth 1 dari website Twitter untuk *developer* dengan cara mendaftarkan aplikasi dan mendapat *customer* dan *tokens*. Peneliti menggunakan API jenis *search* yang memiliki kemampuan untuk mengambil data sesuai *keyword* yang diberikan. Data Twitter yang terambil memiliki format JSON dan masih berupa data yang tidak terstruktur. Atribut yang diambil hanya atribut *text*, *created_at* dan *id_user*. Untuk mempermudah penulisan atribut, peneliti mengubah nama atribut tersebut secara berurutan menjadi teks, tanggal dan *id_user*. Peneliti melakukan penambahan atribut untuk mempermudah, atribut yang ditambahkan adalah atribut sentimen dan atribut akun. Atribut sentimen berfungsi untuk menyimpan nilai sentimen yang akan diklasifikasikan, sedangkan atribut akun sebagai penanda bahwa data tersebut berasal dari *keyword* @AhokDjarot dan @JktMajuBersama. Data yang telah terkumpul disimpan menjadi 1 buah fail Json dan pada setiap fail terdapat kurang lebih 4 000 data *tweet*. Data diambil pada tanggal tertentu, yaitu tanggal 4, 12, 16, 17, 18, dan 19 April 2017 di waktu yang berbeda-beda. Tabel 2 menunjukkan hasil data yang diambil, dengan jumlah data yang terkumpul pada akun @AhokDjarot adalah 24 377 *tweet*, akun @JktMajuBersama adalah 24 311 *tweet*.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Tabel 2 Data Twitter yang terkumpul

Akun	Tanggal pengambilan data (April 2017)					
	4	12	16	17	18	19
@AhokDjarot	4 097	4 000	4 090	4 072	4 052	4 066
@JktMajuBersama	4 000	4 020	4 081	4 085	4 034	4 091

Data yang terkumpul disimpan ke dalam format JSON. Setiap 1 fail berisikan sekitar 4 000 data yang tersimpan dalam bentuk *list*. Setiap dokumen mengandung 5 atribut beserta dengan nilainya. Tabel 3 menunjukkan 1 dokumen JSON yang tersimpan.

Tabel 3 Contoh *record* data berupa Json

Contoh data json
{ "sentimen": "something", "isi": "RT @NovNuraidah: @aniesbaswedan @sandiuno @JktMajuBersama @Gerindra @PKSejahtera @PartaiPerindo \nSyariah kotanya\nHancur warungnya\nBahagia F\u2026", "tanggal": "Tue Apr 04 14:07:14 +0000 2017", "id_user": 108503037, "akun": "@JktMajuBersama" }

Proses pelabelan data dilakukan oleh satu orang secara manual, yaitu mengambil 600 data secara acak dari kedua akun. Peneliti melakukan pelabelan secara manual dengan cara menentukan secara pribadi suatu data masuk kedalam kelas positif, kelas negatif atau kelas netral. Pengelompokan kelas positif dilihat dari isi *tweet* mengandung kata bermakna positif, mendukung dan pernyataan setuju. Kelas negatif merupakan kelas dengan data yang mengandung kata bermakna negatif, ejekan, dan kontra. Kelas netral merupakan kelas yang berisikan data mengandung berita. Tabel 4 menunjukkan contoh data yang dilabelkan secara manual.

Tabel 4 Hasil pelabelan data secara manual

Label	Isi	Pertimbangan	Kata Kunci
Positif	@AhokDjarot Melayani warga jakarta mulai dari lahir dengan berbagai program.. #FreeAhok @basuki_btp https://t.co/jrzDjWO5EE	Kata bermakna positif	Melayani
Positif	RT @RahyaMaya: https://t.co/pvVUIp3YYw Pak Ahok itu didzolimi... Aku yakin Pak Ahok bebas... #FreeAhok @basuki_btp @AhokDjarot https://t.co/u2026	Mendukung	Yakin dan bebas

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
- Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Tabel 4 Lanjutan

Label	Isi	Pertimbangan	Kata Kunci
Positif	RT @Jakarta_Kece: Pak Ahok Djarot paling kece badai yang bisa menata kota Jakarta... #FreeAhok @basuki_btp @AhokDjarot @ezkisuyanto @mrshan	Mendukung	Kece dan bisa
Negatif	@Lintank01 @AhokDjarot Golongan sumbu pendek lo. Fpi piaraan kluarga cendana. Bibib risiek ulama mesum. Bisanya demo gak mau kerja.	Ejekan	Golongan sumbu pendek, piaraan dan mesum
Negatif	@MudasirRomini @AhokDjarot lah ahok di penjara.Bersih2 penjara aja biar dekat dgn ahok	Kontra	Di penjara
Negatif	@Jakarta_Kece @Fakta @AhokDjarot Wkwkw....ada udang dbalik batu itu.	Kata bermakna negatif	Udang di balik batu
Netral	RT @AhokDjarot: Penasaran sama #BasukiDjarot? Punya #PertanyaanKepo yg ingin dijawab? RT dgn #KepoinPelayanJakarta ! Pertanyaan terpilih	Berita	Pertanyaan terpilih
Netral	@KompasTV @Rosianna766Hi @basuki_btp @AhokDjarot #AhokDjarotDiRosi Sebagai warga diluar DKI berharap Pilkada DKI Pu\u2026 https://t.co/hce2ZPDsln	Berita	Warga dan Pilkada DKI
Netral	@AhokDjarot Pak ahok saya tkw saudi. saya ingin kerja bersih bersih masjid, Nanti 1 tahun lagi contract selesai.	Berita	Saya tkw, kerja dan kontrak

Pembagian Data

Setelah peneliti melakukan pelabelan 600 data dari gabungan kedua akun kemudian data tersebut dibagi menjadi data latih dan data uji secara acak. Pembagian data latih terdiri dari 300 data akun @AhokDjarot dan 100 data akun @JktMajuBersama. Data uji terdiri dari 100 data dari setiap akun.

Praproses Data

Tahap praproses diawali dengan pembersihan data, setelah itu normalisasi data, *stemming data*, dan penghapusan *stopword*. Sebelum itu, semua data dipisah setiap kata.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Pembersihan Data

Pembersihan data dilakukan dengan cara melakukan penghapusan terhadap kata atau karakter yang mengurangi kualitas data latih. Pertama adalah mencari kata dengan menggunakan fungsi ekspresi reguler yang ada di Python, yaitu mencari kata yang berisikan *link* dengan diawali www atau http, mencari akun Twitter yang ada di dalam data dengan awalan @ dan mencari kata yang diawali *hashtag*. Setelah itu, setiap kata yang ditemukan, langsung dihapus.

Tabel 5 menggambarkan hasil dari proses penghapusan data.

Tabel 5 Hasil pembersihan data

No	Sebelum Pembersihan Data	Sesudah Pembersihan Data
1	rt @ahokdjarot: penasaran sama #basukidjarot? punya #pertanyaankepo yg ingin dijawab? rt dgn ! #kepoinpelayanjakarta ! pertanyaan terpilih ak...	rt penasaran sama punya yg ingin dijawab? rt dgn ! pertanyaan terpilih ak...
2	@ahokdjarot melayani warga jakarta mulai dari lahir dengan berbagai program.. #freeahok @basuki_btp https://t.co/jrzdjwo5ee	melayani warga jakarta mulai dari lahir dengan berbagai program..

Normalisasi Data

Tahap normalisasi mengubah kata kata yang tidak baku menjadi kata yang lebih baku. Selain itu juga dilakukan penghapusan terhadap karakter yang diulang (Aziz 2013). Data normalisasi yang digunakan sejumlah 3720 pasang kata berdasarkan penelitian Aziz (2013). Setiap pasang kata yang ada di data tersebut terdiri dari kata tidak baku dan pasangan kata bakunya. Program dijalankan dengan cara setiap menemukan kata tidak baku di dalam data *tweet* maka akan diganti dengan kata baku yang ada. Tabel 6 adalah contoh hasil normalisasi data.

Tabel 6 Contoh hasil normalisasi

No	Sebelum Normalisasi	Sesudah Normalisasi
1	rt penasaran sama punya yg ingin dijawab? rt dgn ! pertanyaan terpilih ak...	rt penasaran sama punya yang ingin dijawab? rt dengan ! pertanyaan terpilih ak...
2	melayani warga jakarta mulai dari lahir dengan berbagai program..	melayani warga jakarta mulai dari lahir dengan berbagai program..

Stemming Data

Pengubahan kata yang mengandung imbuhan menjadi kata dasar. Imbuhan ini termasuk imbuhan yang ada di depan, di antara, atau di akhir kata tersebut. Sastrawi memiliki 40 aturan pemenggalan yang digunakan untuk melakukan proses stemming. Selama proses stemming, juga dilakukan proses penghapusan pungtuasi. Penghapusan dilakukan dengan cara melakukan pengecekan terhadap setiap karakter, kemudian apabila karakter tersebut merupakan pungtuasi, maka akan dihapuskan. Tabel 7 merupakan contoh hasil proses stemming data dan penghapusan pungtuasi.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
- Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Tabel 7 Contoh hasil *stemming data* dan penghapusan pungtuasi

No	Sebelum Stemming	Setelah Stemming
1	rt penasaran sama punya yang ingin dijawab? rt dengan ! pertanyaan terpilih ak...	rt penasaran sama punya yang ingin jawab rt dengan tanya pilih ak
2	melayani warga jakarta mulai dari lahir dengan berbagai program..	layan warga jakarta mulai dari lahir dengan bagai program

Penghapusan Stopword

Total data *stopword* yang digunakan adalah 762 kata berdasarkan Tala (2003) dan peneliti menambahkan kata rt ke dalam *stopword*. Setiap kata yang ada di data diperiksa apakah terdapat kata yang sama dengan *stoplist*, jika ada, maka kata tersebut dihapuskan. *Stoplist* penelitian ini menggunakan hasil analisis frekuensi kata tertinggi yang dilakukan oleh Tala (2003). Tabel 8 adalah contoh dari data yang telah melewati proses penghapusan *stopword*.

Tabel 8 Contoh hasil penghapusan *stopword*

No	Sebelum Penghapusan <i>Stopword</i>	Setelah Penghapusan <i>Stopword</i>
1	rt penasaran sama punya yang ingin jawab rt dengan tanya pilih ak	penasaran pilih ak
2	layan warga jakarta mulai dari lahir dengan bagai program	layan warga jakarta lahir program

Hasil praproses direpresentasikan melalui *wordcloud* yang dapat dilihat pada Gambar 3. Visualisasi *wordcloud* dilakukan pada setiap kelas berdasarkan data latih yang telah melalui praproses data.



Gambar 3 *Wordcloud* dari hasil praproses setiap kelas (a) Kelas positif (b) Kelas negatif (c) Kelas netral

Pemodelan Klasifikasi

Data bersih yang didapatkan dari hasil praproses kemudian dilakukan pemisahan pada setiap kata. Kemudian dihitung frekuensi dari setiap kata tersebut pada semua dokumen. Hasilnya dijadikan sebagai *word list* dan dilakukan ekstraksi fitur kepada hasil tersebut. Ekstraksi fitur didapat dari perhitungan kemunculan setiap kata pada setiap data di dalam dokumen kemudian dijadikan sebagai *training*

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajar IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

Most Informative Features		
contains(pilkada) = True	netral : positi =	9.8 : 1.0
contains(hadiah) = True	netral : positi =	9.8 : 1.0
contains(nya) = True	negati : positi =	9.7 : 1.0
contains(tonton) = True	netral : positi =	8.0 : 1.0
contains(annies-sandi) = True	netral : positi =	8.0 : 1.0
contains(damai) = True	netral : positi =	8.0 : 1.0
contains(nonton) = True	netral : positi =	8.0 : 1.0
contains(jujur) = True	negati : positi =	6.6 : 1.0

Gambar 4 Model *classifier* yang paling berpengaruh untuk menentukan kelas

Pengujian

Model *classifier* digunakan sebagai penentu untuk pengujian data menggunakan data uji. Setiap data diprediksi berdasarkan 3 kelas, yaitu positif, negatif, atau netral. Data uji yang digunakan berjumlah 200 data. Setiap kata yang telah dipisah, kemudian akan diprediksi menggunakan model yang sudah ada. Proses prediksi kelas yang dilakukan, menggunakan *library* NLTK berdasarkan metode Naïve Bayes. Contoh hasil dari pengujian model bisa dilihat pada Tabel 9.

Tabel 9 Hasil pengujian model

Isi	Sentimen Awal	Prediksi Model
@AhokDjarot Melayani warga jakarta mulai dari lahir dengan berbagai program.. #FreeAhok @basuki_btp https://t.co/jrzDjWO5EE	Positif	Positif
@Lintank01 @AhokDjarot Golongan sumbu pendek lo. Fpi piaraan kluarga cendana. Bibib risiek ulama mesum. Bisanya demo gak mau kerja.	Negatif	Negatif
RT @AhokDjarot: Penasaran sama Netral #BasukiDjarot? Punya #PertanyaanKepo yg ingin dijawab? RT dgn #KepoinPelayanJakarta ! Pertanyaan terpilih ak\u2026	Netral	Positif

Tabel 9 Lanjutan

Isi	Sentimen Awal	Prediksi Model
Berapa banyak ya kembarannya pak wagub? @MetroTVToday @basuki_btp @AhokDjarot @PartaiSocmed @kurawa	Netral	Negatif

Evaluasi Hasil

Proses evaluasi dari program dilakukan dengan menggunakan *confusion matrix*. Untuk setiap kelas, dilihat berapa banyak yang *true positif* dan *false negatif*. Proses evaluasi menggunakan data latih berjumlah 400 data dan data uji berjumlah 200 data. Dari proses pengujian data uji didapatkan hasil seperti pada Tabel 10.

Tabel 10 Hasil *confusion matrix table*

	Kelas	Prediksi		
		Positif (%)	Negatif (%)	Netral (%)
Aktual	Positif	85.29	8.82	5.88
	Negatif	31.15	68.85	0
	Netral	59.42	11.59	28.99

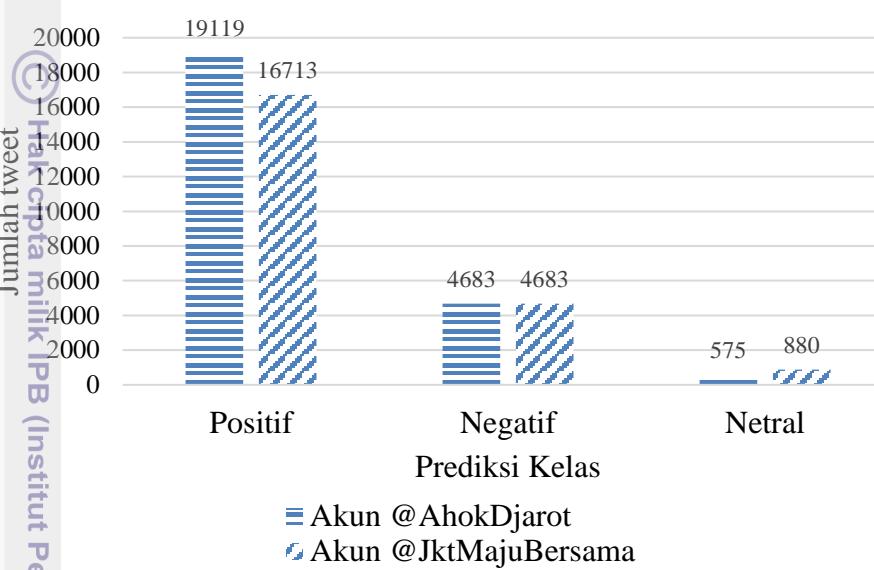
Berdasarkan Tabel 10, kemudian dihitung untuk tingkat akurasi dari model yang dibuat, data uji memiliki tingkat akurasi sebesar 60.60% dengan menggunakan data latih sebanyak 400 data. Selain menggunakan tingkat akurasi, proses evaluasi model juga dilihat dari tingkat sensitifitas dan spesifisitas setiap kelas dalam mengelompokkan suatu data. Hasil perhitungan dapat dilihat di Tabel 11.

Tabel 11 Hasil sensitifitas dan spesifisitas setiap kelas

Label	Sensitifitas (%)	Spesifititas (%)
Positif	85.29	53.94
Negatif	68.85	89.78
Netral	28.98	96.89

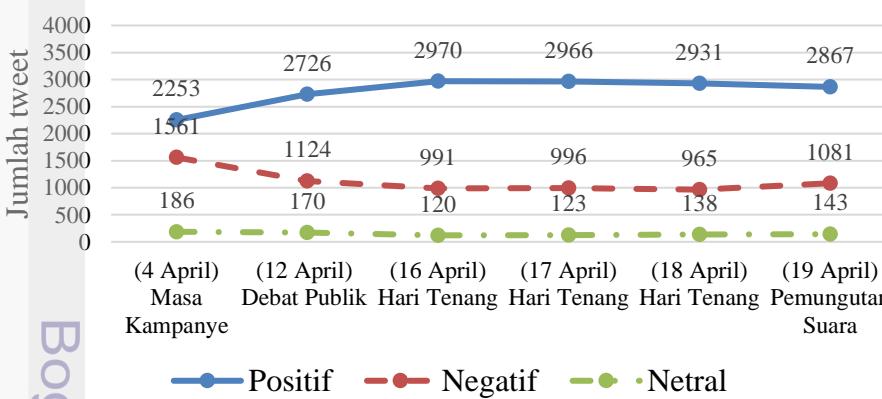
Berdasarkan Tabel 11 dapat dilihat bahwa kelas positif memiliki sensitifitas yang lebih tinggi dari pada kelas negatif dan netral. Namun, kelas positif memiliki spesifisitas yang paling kecil dibandingkan dengan kelas yang lainnya. Kelas positif memiliki sensitifitas yang tinggi dikarenakan data dikelas positif lebih terlihat perbedaannya daripada kelas yang lain. Kelas netral, memiliki nilai sensitifitas yang paling rendah karena data pada kelas netral sulit dibedakan dengan kelas positif. Hal ini berkaitan dengan tingkat spesifitas kelas positif yang rendah. Berarti, model mengalami kesulitan dalam memprediksi kelas yang bukan positif dan kelas netral. Faktor lain yang mempengaruhi hasil adalah dilakukannya pelabelan data hanya oleh satu orang, sehingga menjadikan proses pelabelan masih bersifat subjektif.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Gambar 5 Hasil Total Prediksi Sentimen

Gambar 6 menunjukkan jumlah sentimen positif, negatif dan netral dari pasangan cagub dan cawagub nomor 3. Berdasarkan Gambar tersebut, dapat disimpulkan bahwa data yang terprediksi positif relatif mengalami peningkatan dan mengalami penurunan pada kelas negatif kecuali pada tanggal 19 April, sedangkan kelas netral relatif simbang.



Gambar 6 Hasil sentimen pasangan calon nomor 3 @JktMajuBersama

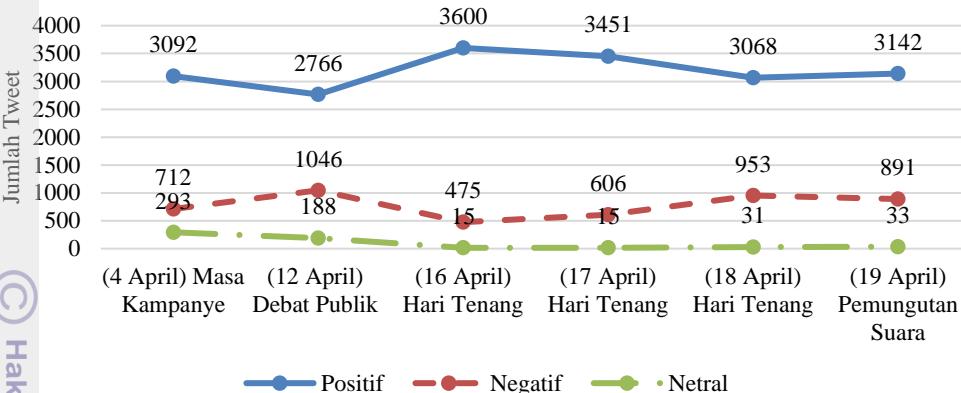
Gambar 7 menjelaskan bahwa hasil prediksi paslon 2 pada kelas positif mengalami penurunan hingga tanggal 16 April 2017 dan pada tanggal 17 April

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

- Hak Cipta Dilindungi Undang-Undang**
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.

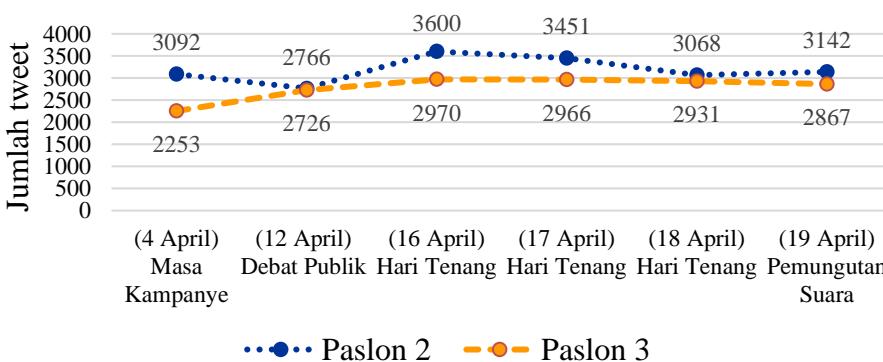
© Hak cipta milik IPB (Institut Pertanian Bogor)

2017 mengalami kenaikan, namun hal ini berkebalikan dengan kelas negatif. Kelas netral memiliki jumlah data paling banyak pada tanggal 16 April 2017 dan paling rendah pada tanggal 12 April 2017.



Gambar 7 Hasil Sentimen Pasangan Nomor 2 @AhokDjarot

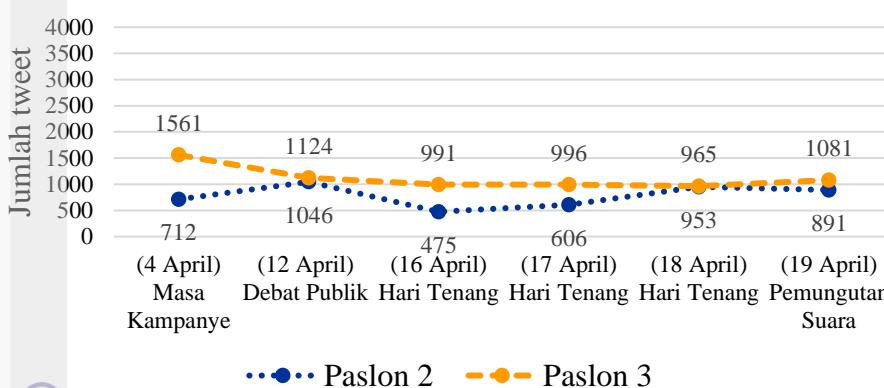
Selain melihat hasil sentimen setiap pasangan calon, peneliti juga melakukan perbandingan berdasarkan kelas yang ada. Hasil tersebut didapat berdasarkan prediksi pada data yang telah dikumpulkan pada akun @AhokDjarot sebanyak 24 377 tweet, akun @JktMajuBersama sebanyak 24 311 tweet. Gambar 8 menunjukkan perbandingan hasil prediksi sentimen positif dari pasangan cagub dan cawagub. Paslon nomor 2 memiliki hasil lebih banyak pada paslon nomor 3 sentimen positif hampir di seluruh tanggal kecuali tanggal 12. Paslon nomor 3 memiliki hasil yang relatif bertambah, namun terjadi penurunan mulai tanggal 17 April 2017.



Gambar 8 Perbandingan kelas positif dari 2 akun pasangan calon

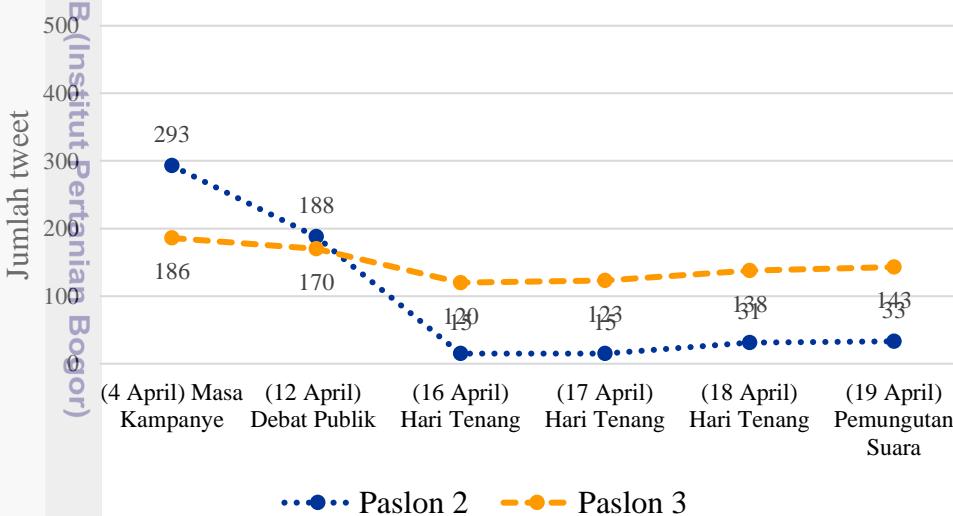
Prediksi sentimen negatif pada paslon nomor 3 cenderung menurun mulai dari tanggal 4 April 2017 dan ada sedikit peningkatan pada tanggal 19 April 2017. Paslon nomor 2 memiliki jumlah prediksi tertinggi pada tanggal 12 April dan terendah pada tanggal 16 April 2017. Hasil prediksi kedua paslon terjadi perbedaan yang sangat tipis di tanggal 18 April 2017 dengan selisih 50 data. Penjelasan grafik dapat dilihat pada Gambar 9.

- Hak Cipta Dilindungi Undang-Undang**
- © Hak Cipta termasuk IPB (Institut Pertanian Bogor)
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



Gambar 9 Perbandingan kelas negatif dari 2 akun pasangan calon

Hasil dari predksi sentimen netral pada paslon nomor 2 mengalami penurunan paling rendah pada tanggal 16 April 2017, sedangkan pada paslon nomor 3 pada tanggal yang sama hanya mengalami sedikit penurunan. Paslon nomor 3 memiliki jumlah sentimen netral yang lebih banyak dari pada paslon nomor 2. Gambar 10 menunjukkan perbandingan dari hasil sentimen netral paslon nomor 2 dan paslon nomor 3.



Gambar 10 Perbandingan kelas netral dari 2 akun pasang calon

Informasi berdasarkan website KPU, hasil perhitungan suara dari 13 034 TPS dan 5 591 577 suara yang terkumpul didapatkan paslon nomor 3 yaitu Anies dan Sandi menjadi Gubernur dan Wakil Gubernur DKI Jakarta 2017. Hasil suara yang diperoleh 57.95% untuk pasangan Anies-Sandi dan 42.05% untuk pasangan Ahok-Djarot. Jika dilihat pada predksi sentimen pada penelitian ini, pasangan Ahok-Djarot memiliki keunggulan di sentimen positif dan sentimen negatif lebih rendah dari pasangan Anies-Sandi. Hasil tersebut berbanding terbalik dengan hasil dari pemilu, hal ini bisa disebabkan karena model pada penelitian kurang bisa memprediksi sentimen secara baik. Komposisi data latih yang terlalu sedikit dibanding dengan data yang diprediksi, hal ini mengakibatkan model kurang optimal, terbukti dengan akurasi yang rendah dan spesifitas dan sensitifitas yang kecil.



SIMPULAN DAN SARAN

Simpulan

Pengambilan sampel data Twitter menggunakan Twitter API dilakukan oleh peneliti selama tanggal 4, 12, 16, 17, 18, dan 19 April 2017. Data yang diambil menggunakan *keyword* @AhokDjarot dan @JktMajubersama. Model yang dibuat telah berhasil memprediksi setiap data pada data uji dengan akurasi 60.60% menggunakan metode Naïve Bayes. Sensitifitas terbesar pada kelas positif, kemudian kelas negatif dan kelas netral. Spesifisitas terbesar pada kelas netral, kemudian kelas negatif dan kelas positif.

Saran

Proses pelabelan data menjadi dasar untuk membangun sebuah model yang baik. Penelitian ini masih melakukan pelabelan oleh 1 orang, sehingga hasil dari pelabelan masih bersifat subjektif, sehingga perlu dilakukan proses pelabelan yang melibatkan lebih dari 1 orang agar hasil pelabelan data lebih objektif. Hal lain yang menjadi dasar dalam pembuatan model adalah jumlah data latih yang tersedia. Penelitian ini memiliki data latih yang sedikit. Perlu dilakukan penambahan data latih untuk melihat perubahan hasil prediksi untuk setiap penambahan data latih.

DAFTAR PUSTAKA

- Asian J. 2007. Effective Techniques for Indonesian Text Retrieval [tesis]. Melbourne(AU): University Australia.
- Aziz M. 2013. Sistem Pengklasifikasian Entitas pada Pesan Twitter menggunakan Ekspresi Regular dan Naive Bayes [skripsi]. Bogor(ID): Institut Pertanian Bogor.
- Han J, Kamber M. 2011. *Data mining: concepts and techniques*. 3rd ed. Burlington, MA: Elsevier.
- Mahendra, Oka AA. 2012. Kampanye Pemilu 2014 sebagai bagian dari Pendidikan Politik Masyarakat. *Jurnal Legislasi Indonesia*. 9(4):547-562.
- Pak A, Paroubek P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Di dalam: *LREC*. [internet] Vol. 10. [diunduh 2017 Agu 29]. Tersedia pada: <http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf>
- Schrape JF. 2016. Social Media, Mass Media and the Public Sphere. Differentiation, Complementarity and Co-Existence. [diunduh 2017 Agu 29]. Tersedia pada: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2858891
- Situmorang JR. 2012. Pemanfaatan internet sebagai new media dalam bidang politik, bisnis, pendidikan, dan sosial budaya. *Jurnal Administrasi Bisnis*. 8(1):73-87.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



[Social Memos] Social Media Statistic for Indonesia. [Internet]. [diunduh 2017 Agustus 16]. Tersedia pada: <http://socialmemos.com/social-media-statistics-for-indonesia/>

[Statista] The Statistic Portal. 2017. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017. [Internet]. [diunduh 2017 Agustus 3]. Tersedia pada : <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>

Stieglitz S, Xuan LD. 2012. Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*. 3(16): 1277–1291. doi:10.1007/s13278-012-0079-3

Tala FZ. 2003. A study of stemming effects on information retrieval in Bahasa Indonesia. *Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands*. [diunduh 2017 Agu 24]. Tersedia pada: <http://ai2-s2-dfs.s3.amazonaws.com/8ed9/c7d54fd3f0b1ce3815b2eca82147b771ca8f.pdf>

Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. Di dalam: *Proceedings of the ACL 2012 System Demonstrations*. [internet] Association for Computational Linguistics. hlm. 115–120. [diunduh 2017 Agu 24]. Tersedia pada: <http://dl.acm.org/citation.cfm?id=2390490>

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.



RIWAYAT HIDUP

Penulis lahir pada tanggal 6 November 1994 sebagai anak kedua dari pasangan Suranto SE dan Dra Sri Herlina. Penulis memiliki seorang kakak bernama Nurul Hidayah SSi. Penulis menamatkan masa belajar jenjang sekolah menengah di SMA Negeri 1 Kendal. Pada tahun 2013 melanjutkan ke jenjang perguruan tinggi di Departemen Ilmu Komputer IPB melalui jalur SNMPTN.

Perkuliahannya tahun pertama, penulis menjadi RT di Lorong 3 C3 Asrama IPB dan menjadi ketua divisi Dekorasi dan Dokumentasi pada kepanitian ISEE (*International Scholarship Education and Expo*). Selain itu penulis mengikuti Klub Fotografi IPB Shutter dan Organisasi Mahasiswa Daerah Kendal bernama Fokma Bahurekso Kendal. Penulis mengikuti Lembaga Da'wah Fakultas pada tahun ke tiga sebagai ketua Creative Media. Tahun 2016 penulis mengikuti kegiatan praktik kerja lapang di Fujitsu Indonesia.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.

b. Pengutipan tidak merugikan kepentingan yang wajib IPB.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB.