SISTEM ANALISIS SENTIMEN PUBLIK TENTANG OPINI PEMILIHAN KEPALA DAERAH JAWA TIMUR 2018 PADA DOKUMEN TWITTER MENGGUNAKAN *NAIVE BAYES CLASSIFIER*

SKRIPSI

OLEH
FAJAR DARWIS DZIKRIL HAKIMI
NIM. H02214001

PROGRAM STUDI MATEMATIKA
JURUSAN SAINS
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
SURABAYA
2018

PERNYATAAN KEASLIAN

Yang bertandatangan di bawah ini:

Nama : Fajar Darwis Dzikril Hakimi

NIM : H02214001

Program Studi : Matematika

Angkatan : 2014

Menyatakan bahwa saya tidak melakukan plagiat dalam bentuk apapun pada skripsi saya yang berjudul: **Sistem Analisis Sentimen Publik Tentang Opini Pemilihan Kepala Daerah Jawa Timur 2018 Pada Dokumen Twitter Menggunakan *Naive Bayes Classifier.*** Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat, maka saya akan menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya

Surabaya, 3 Agustus 2018

Fajar Darwis Dzikril Hakimi

NIM. H02214001

LEMBAR PENGESAHAN

SISTEM ANALISIS SENTIMEN PUBLIK TENTANG OPINI PEMILIHAN KEPALA DAERAH JAWA TIMUR 2018 PADA DOKUMEN TWITTER MENGGUNAKAN *NAIVE BAYES CLASSIFIER*

Disusun oleh
Fajar Darwis Dzikril Hakimi
NIM. H02214001

Telah dipertahankan di depan Dewan Penguji
Pada tanggal 18 Juli 2018
Dan dinyatakan telah memenuhi syarat
untuk memperoleh gelar
Sarjana Matematika (S.Mat)

Dewan Penguji

Penguji I

Ahmad Hanif,Asyhar, M.Si
NIP. 198601232014031001

Penguji II

Nurissaidah Ulinnuha, M. Kom
NIP. 199011022014032004

Penguji III

Aris Fanani, M.Kom
NIP. 198701272014031002

Penguji IV

Wika Dianita Utami, M.Sc
NIP. 1992061002018012003

Mengesahkan
Dekan Fakultas Sains dan Teknologi
UIN Sunan Ampel Surabaya

Dr. Eni Purwati, M.Ag
NIP. 196512211990022001

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI
KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : Fajar Darwis Dzikril Hakimi
NIM : H02214001
Fakultas/Jurusan : Saintek / Matematika
E-mail address : fajarddh@gmail.com

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
☑ Sekripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (.....)
yang berjudul :

Sistem Analisis Sentimen Publik tentang Opini Pemilihan Kepala Daerah Jawa Timur 2018 pada Dokumen Twitter menggunakan Naive Bayes Classifier

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya, 1 Agustus 2018

Penulis

(Fajar Darwis D. H.)

SISTEM ANALISIS SENTIMEN PUBLIK TENTANG OPINI PEMILIHAN KEPALA DAERAH JAWA TIMUR 2018 PADA DOKUMEN TWITTER MENGGUNAKAN *NAIVE BAYES CLASSIFIER*

ABSTRAK

Pada tahun 2018, pemilihan kepala daerah diadakan serentak di sebagian besar daerah di Indonesia. Salah satu daerah tersebut adalah Jawa Timur. Dalam pra-pelaksanaan pemilihan Kepala Daerah Jawa Timur, terdapat berbagai opini masyarakat yang bersentimen positif dan negatif pada twitter. Opini tersebut dapat digunakan sebagai parameter untuk mengukur kekuatan masing-masing calon kepala daerah. Tujuan penelitian ini yaitu untuk mengetahui kecenderungan opini masyarakat tentang pemilihan Kepala Daerah Jawa Timur pada twitter. Dalam penelitian ini dilakukan beberapa tahapan yaitu *crawling* data twitter, pelabelan data, penghapusan data yang tidak dibutuhkan (data *outlier* dan netral), pre-proses data teks, pembuatan sistem klasifikasi dengan *naïve bayes classifier*, dan uji coba sistem pada data twitter yang lebih banyak. Metode pembobotan kata yang digunakan yaitu *Term Frequency* (TF) dan *Term Frequency-Inverse Document Frequency* (TF-IDF). Diantara kedua metode tersebut, pembobotan kata TF mempunyai hasil yang lebih baik. Dari penelitian ini diperoleh hasil performa sistem untuk masing-masing kandidat. Untuk data calon gubernur nomor urut satu didapat hasil akurasi, presisi, *recall*, dan *f-measure* berturut-turut sebagai berikut 98,99%, 93,44%, 97,78%, dan 95,56%. Selanjutnya untuk data calon kepala daerah dengan nomor urut dua diperoleh hasil akurasi, presisi, *recall*, dan *f-measure* berturut-turut sebagai berikut: 98,95%, 97,78%, 98,55%, dan 98,17%. Berdasarkan data yang diperoleh dari twitter, masyarakat Jawa Timur cenderung lebih memilih calon kepala daerah dengan nomor urut satu yaitu Khofifah Indar Parawansa.

Kata Kunci: Analisis Sentimen, Pemilihan Kepala Daerah Jawa Timur, *Naïve Bayes Classifier*

PUBLIC SENTIMENT ANALYSIS SYSTEM TOWARDS EAST JAVA GOVERNOR ELECTION 2018 ON TWITTER USING *NAIVE BAYES CLASSIFIER*

ABSTRACT

In 2018, governor election was held together in most of Indonesian region. One of them was East Java. In pre-implementation of East Java Governor election, there were several public opinions which had positive and negative sentiment on twitter. They could be used as a parameter to evaluate the strength of each candidate. The purpose of this research is to know the tendency of public opinion about East Java Governor election on twitter. In this research, there are several steps to do i.e. crawling data on twitter, labelling the data, deleting the data (outlier and neutral), pre-processing of text data, making classification system using naïve bayes classifier, and applying the system on more twitter data. The word weighting methods used are Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). The result shows that TF word weighting method has better performance. This research has two result of system performance for each candidate. Based on the data of the first candidate, the result of accuracy, precision, recall, and f measure are as follows 98,99%, 93,44%, 97,78%, and 95,56%. Based on the data of the second candidate, the result of accuracy, precision, recall, and f-measure are as follows: 98,95%, 97,78%, 98,55%, and 98,17%. Based on the data from twitter, East Java citizen has tendency to choose the first candidate, Khofifah Indar Parawansa.

Keywords: Sentiment Analysis, East Java Governor Election, Naive Bayes Classifier.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

A. Latar Belakang

Seiring dengan perkembangan zaman, ilmu pengetahuan dan teknologi juga mengalami kemajuan yang signifikan. Banyak aspek kehidupan manusia yang dimudahkan dengan adanya kemajuan ini. Jika pada zaman dahulu manusia melakukan segala sesuatu secara manual, maka sekarang ini sudah banyak hal yang bisa dilakukan secara otomatis. Salah satu contohnya adalah di bidang komunikasi. Pada zaman dahulu, pertukaran informasi biasa dilakukan dengan menggunakan surat. Tapi pada zaman sekarang ini, pertukaran informasi bisa dilakukan dengan hanya mengoperasikan *gadget* di tangan. Itu hanya satu contoh dari sekian banyak contoh kemudahan yang tercipta karena kemajuan teknologi.

Kemajuan ilmu pengetahuan dan teknologi, khususnya di bidang ilmu komputer membuat pertukaran informasi dapat dilakukan dimanapun dan kapanpun (Wongso, dkk., 2017). Hal ini menyebabkan melimpahnya informasi yang beredar di internet. Salah satu cara agar informasi yang melimpah tersebut mempunyai nilai tambah dan dapat dimanfaatkan secara maksimal adalah dengan melakukan klasifikasi informasi. Tentu saja akan dibutuhkan waktu yang lama bagi seseorang yang ingin melakukan pengklasifikasian informasi secara manual. Oleh karena itu, sekarang ini, klasifikasi dokumen secara otomatis banyak dikembangkan karena pengklasifikasian secara manual sudah tidak efektif lagi (Wong & Abednego, 2015).

Pada zaman yang modern ini, media sosial mempunyai peranan penting sebagai media kampanye pemilu. Hal ini dikarenakan media sosial memberikan kemudahan kepada penggunanya untuk mengakses berita politik terbaru dan opini masyarakat yang tidak dimuat dalam media cetak ataupun media penyiaran (Lestari, Perdana, & Fauzi, 2017). Selain itu, media sosial juga memudahkan penggunanya untuk menyampaikan opini mereka tentang berita politik terbaru. Salah satu media sosial yang paling sering digunakan untuk penyampaian opini ataupun media kampanye adalah twitter. Hal ini dikarenakan kelebihan twitter yang mudah diakses, jumlah pengikut tidak terbatas, dan jumlah karakter yang digunakan hanya 280 karakter sehingga mampu menyampaikan maksud dengan singkat, padat, dan jelas.

Dalam menyampaikan opini dalam media sosial, tentu akan lebih baik jika dilakukan dengan memperhatikan aturan yang ada dalam syariat Islam. Dalam Islam, ada beberapa aturan yang harus diperhatikan ketika ingin menyampaikan pendapat kepada orang lain. Salah satu aturan tersebut adalah harus menginformasikan kebenaran berdasarkan fakta dan juga tidak merekayasa atau memanipulasi fakta. Hal ini dijelaskan dalam Qur'an Surat Al-Hajj ayat 30 berikut:

ذَٰلِكَ ۖ وَمَن يُعَظِّمْ حُرُمَٰتِ ٱللَّهِ فَهُوَ خَيْرٌ لَّهُۥ عِندَ رَبِّهِۦ ۗ وَأُحِلَّتْ لَكُمُ ٱلْأَنْعَٰمُ إِلَّا مَا يُتْلَىٰ عَلَيْكُمْ ۖ فَٱجْتَنِبُوا۟ ٱلرِّجْسَ مِنَ ٱلْأَوْثَٰنِ وَٱجْتَنِبُوا۟ قَوْلَ ٱلزُّورِ ﴿٣٠﴾

"30. Demikianlah (perintah Allah). Dan barangsiapa mengagungkan apa-apa yang terhormat di sisi Allah maka itu adalah lebih baik baginya di sisi Tuhannya. Dan telah dihalalkan bagi kamu semua binatang ternak, terkecuali yang diterangkan kepadamu keharamannya, maka jauhilah olehmu berhala-berhala yang najis itu dan jauhilah perkataan-perkataan dusta."

Selain itu, dalam menyampaikan pendapat juga harus dilakukan secara bijaksana dan tidak menyakiti hati orang lain. Hal ini seperti dijelaskan dalam Qur'an Surat an-Nisa ayat 9 berikut ini:

وَلْيَخْشَ ٱلَّذِينَ لَوْ تَرَكُوا۟ مِنْ خَلْفِهِمْ ذُرِّيَّةً ضِعَٰفًا خَافُوا۟ عَلَيْهِمْ فَلْيَتَّقُوا۟ ٱللَّهَ وَلْيَقُولُوا۟ قَوْلًا سَدِيدًا ﴿٩﴾

"9. Dan hendaklah takut kepada Allah orang-orang yang seandainya meninggalkan dibelakang mereka anak-anak yang lemah, yang mereka khawatir terhadap (kesejahteraan) mereka. Oleh sebab itu hendaklah mereka bertakwa kepada Allah dan hendaklah mereka mengucapkan perkataan yang benar qaulan sadida."

Qaulan sadida dalam ayat 9 surat an-Nisa di atas ditafsirkan sebagai perkataan yang benar dan bijaksana. Oleh karena itu, dalam menyampaikan pendapat dalam kehidupan sehari-hari, sebaiknya dilakukan dengan berdasarkan fakta yang ada dan disampaikan secara bijaksana. Hal ini juga berlaku untuk penyampaian pendapat di media sosial. Ketika seseorang membuat pendapat di media sosial, tentu pendapat tersebut berpotensi untuk dibaca oleh banyak orang. Oleh karena itu, ketika menyampaikan pendapat, sebaiknya dilakukan dengan berdasarkan fakta dan juga tidak menyinggung pihak atau golongan tertentu.

Tahun 2018 merupakan tahun diadakannya pemilihan kepala daerah serentak hampir di seluruh daerah di Indonesia. Jawa Timur merupakan salah satu daerah yang ikut ambil bagian pada pemilihan kepala daerah serentak pada tahun 2018 ini. Pada pemilihan Kepala Daerah Jawa Timur, terdapat dua calon pasangan gubernur dan wakil gubernur. Dalam pra-pelaksanaan pemilihan Kepala Daerah Jawa Timur, terdapat berbagai opini masyarakat dengan sentimen positif dan negatif pada

dokumen twitter. Opini masyarakat tersebut bisa dijadikan peta kekuatan calon gubernur dan wakil gubernur dalam persiapan menghadapi pemilu. Oleh karena itu, dengan melakukan klasifikasi terhadap dokumen twitter maka peta kekuatan calon gubernur dan wakil gubernur akan diketahui sedini mungkin sehingga dapat dijadikan evaluasi terhadap kinerja kampanye masing-masing calon.

Masalah akan muncul ketika dilakukan analisis dan pengklasifikasian sentimen secara manual. Masalah tersebut yaitu dibutuhkan waktu yang lama untuk melakukan klasifikasi terhadap dokumen twitter tersebut dan hal tersebut tentu tidak efektif mengingat proses evaluasi terhadap kinerja kampanye harus dilakukan secepat dan seefektif mungkin. Untuk memecahkan permasalahan tersebut, akan digunakan solusi berupa sistem analisis sentimen positif dan negatif tentang opini pemilihan Kepala Daerah Jawa timur 2018 pada dokumen twitter.

Untuk mengklasifikasikan dokumen teks, algoritma yang sering digunakan dan mempunyai hasil yang cukup bagus adalah *naive bayes classifier* dan *support vector machine* (Aggarwal & Zhai, 2013). Selain itu, pada penelitian yang dilakukan oleh Ahmad Fathan Hidayatullah dan Azhari SN (Hidayatullah & SN, 2014) yang berjudul "Analisis Sentimen dan Klasifikasi Kategori Terhadap Tokoh Publik pada Twitter" digunakan data *tweet* sebanyak 1329 data yang diperoleh dari hasil *crawling* yang diberi label kategori secara manual untuk kemudian diklasifikasikan dengan metode *naïve bayes*. Proses yang pertama dilakukan adalah tahap pre-proses untuk membersihkan data *tweet* dan mempersiapkannya untuk proses klasifikasi. Proses dilanjutkan dengan menghitung probabilitas masing-masing kata dalam *tweet* berdasarkan data *training* menggunakan metode *term*

frequency dan *term frequency-inverse document frequency*. Dari dua metode pembobotan kata tersebut, didapat akurasi yang hampir sama dimana *term frequency* mempunyai akurasi sebesar 79,91% sedangkan *term frequency-inverse document frequency* mempunyai akurasi sebesar 79,68%. Ghulam Asrofi Buntoro juga melakukan penelitian yang sejenis (Buntoro, 2016) yang berjudul "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter" digunakan metode *Naïve Bayes Classifier* dan *Support Vector Machine* untuk melakukan klasifikasi terhadap data twitter yang berjumlah 300 data. Hasil terbaik didapat ketika digunakan metode *Naïve Bayes Classifier* dengan nilai akurasi sebesar 95%. Oleh karena itu, dalam penelitian ini digunakan metode *naive bayes classifier* dan pembobotan kata TF dan TF-IDF untuk membuat sistem klasifikasi sentimen positif dan negatif tentang opini pemilihan Kepala Daerah Jawa Timur 2018 pada dokumen twitter.

B. Rumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan di atas, rumusan masalah yang ada di dalam penelitian kali ini adalah:

1. Bagaimana penerapan *naive bayes classifier* untuk mengklasifikasikan sentimen dokumen twitter pada kasus opini pemilihan Kepala Daerah Jawa Timur 2018?
2. Bagaimanakah kecenderungan masyarakat dalam pemilihan Kepala Daerah Jawa Timur 2018 berdasarkan dokumen twitter?

C. Tujuan Penelitian

Penelitian ini dilakukan dengan tujuan yaitu:

1. Mengetahui penerapan *naive bayes classifier* untuk mengklasifikasikan sentimen dokumen twitter pada kasus opini pemilihan Kepala Daerah Jawa Timur 2018.
2. Mengetahui kecenderungan masyarakat dalam pemilihan Kepala Daerah Jawa Timur 2018 berdasarkan dokumen twitter.

D. Batasan Masalah

Dalam penelitian ini, terdapat beberapa batasan masalah yaitu sebagai berikut:

1. Metode yang digunakan adalah *naive bayes classifier*.
2. Data yang digunakan yaitu data berupa teks opini masyarakat yang diambil dari twitter.
3. Output klasifikasi dibagi ke dalam dua kategori sentimen yaitu sentimen positif dan negatif.

E. Manfaat Penelitian

Hasil penelitian ini diharapkan mempunyai manfaat sebagai berikut:

1. Untuk menambah wawasan dan pengetahuan tentang implementasi *naive bayes classifier*.
2. Sebagai bahan rujukan dalam penelitian mengenai klasifikasi teks.
3. Penelitian ini juga diharapkan dapat digunakan sebagai pemetaan kekuatan calon gubernur dan wakil gubernur Jawa Timur sehingga dapat dijadikan evaluasi kinerja kampanye masing-masing calon.

F. Sistematika Penulisan

Penulisan dalam skripsi ini disusun dengan sistematika yang terdiri dari beberapa bab seperti berikut ini:

1. Bab I Pendahuluan, berisi tentang latar belakang masalah, rumusan masalah, tujuan penelitian, batasan masalah, manfaat penelitian, dan sistematika penulisan.
2. Bab II Tinjauan Pustaka, berisi tentang penelitian sebelumnya yang mendasari penelitian ini, teori-teori yang menjadi dasar penelitian ini, dan teori-teori yang digunakan untuk menyelesaikan permasalahan pada penelitian ini.
3. Bab III Metode Penelitian, berisi tentang jenis penelitian, data yang digunakan, dan teknis analisis data.
4. Bab IV Hasil dan Pembahasan, berisi tentang hasil yang didapat dari penelitian ini proses demi proses dan hasil klasifikasi sentimen menggunakan *naive bayes classifier*.
5. Bab V Penutup, berisi tentang kesimpulan yang di dapat dari penelitian ini serta saran untuk perbaikan penelitian selanjutnya.

BAB II

TINJAUAN PUSTAKA

Pada bab ini, dijelaskan tentang teori-teori yang mendasari penelitian ini dan teori-teori yang digunakan untuk menyelesaikan permasalahan yang ada dalam penelitian ini.

A. Analisis Sentimen

Menurut Kamus Besar Bahasa Indonesia (KBBI) *online*, sentimen adalah:

1. Pendapat atau pandangan yang didasarkan pada perasaan yang berlebihan-lebihan terhadap sesuatu (bertentangan dengan pertimbangan pikiran). Contoh: keputusan yang dihasilkan akan tidak adil jika disertai rasa sentimen pribadi.
2. Emosi yang berlebihan. Contoh: rasa sentimen sebagai bangsa Indonesia akan tumbuh kuat jika kita jauh dari negeri ini.
3. Iri hari; tidak senang; dendam.
4. Reaksi yang tidak menguntungkan. Contoh: penurunan harga saham hanya disebabkan oleh sentimen pasar.

Menurut Liu, analisis sentimen atau sering disebut juga dengan opinion mining adalah studi komputasi untuk mengenali dan mengekspresikan opini, sentimen, evaluasi, sikap, emosi, subjektifitas, penilaian atau pandangan yang terdapat pada suatu teks (Liu, 2012).

Sentimen atau opini masyarakat merupakan hal yang penting bagi hampir semua aspek kehidupan manusia. Dalam dunia bisnis dan industri, pihak produsen atau penyedia jasa akan selalu memperhatikan opini masyarakat tentang produk

yang mereka pasarkan. Dalam pemerintahan, pemerintah tentu akan memperhatikan opini masyarakat tentang kebijakan-kebijakan yang sudah ditetapkan. Opini masyarakat ini tentu berguna untuk memperbaiki dan menyempurnakan produk atau kebijakan yang telah dibuat.

Dengan pertumbuhan media sosial yang pesat saat ini seperti twitter, facebook, blog, beberapa produsen dan penyedia jasa menggunakan opini di media sosial tersebut untuk pertimbangan dalam membuat kebijakan di masa depan (Liu, 2012).

Selain itu, opini atau sentimen di media sosial juga bisa digunakan sebagai pemetaan kekuatan calon kepala daerah. Dengan memperhatikan sentimen positif dan negatif yang ditujukan kepada calon kepala daerah tertentu, maka bisa dilihat kecenderungan masyarakat suatu daerah dalam menentukan pilihan dalam pemilihan kepala daerah selanjutnya.

B. Kategorisasi Sentimen Opini

Dalam penelitian sebelumnya yang dilakukan oleh Faradhillah (Faradhillah, 2016), kategorisasi sentimen dilakukan secara manual. Kategorisasi sentimen ini dilakukan agar data yang dimiliki mempunyai kategori sehingga bisa dilakukan proses pelatihan dan pengujian data. Dalam melakukan kategorisasi secara manual, data yang dimiliki dibagi menjadi dua jenis data yaitu data untuk calon gubernur nomor urut satu dan data untuk calon gubernur nomor urut dua. Pembagian data ini dilakukan agar tidak terjadi bias dalam proses kategorisasi opini. Jadi, ketika data itu sudah terbagi dan masuk dalam data untuk salah satu calon, maka tidak akan

terjadi bias dalam pengkategorian opini karena pandangan tentang sentimen tersebut hanya dilihat dari satu sudut pandang saja.

Tabel 2.1 menampilkan beberapa contoh opini masyarakat tentang tokoh Khofifah Indar Parawansa yang mempunyai sentimen positif, negatif, dan netral tentang pemilihan kepala daerah:

Tabel 2.1 Contoh Pengkategorian Opini tentang Khofifah

Opini	Klasifikasi Opini
Gus Sholah menyebut duet Khofifah-Emil sangat klop dan pas.	Positif
Khofifah merupakan pemimpin wanita yang cerdas, jujur, amanah, dan brilian. No 1 Pasti Bisa !	Positif
Bu Khofifah belum berpengalaman memimpin Jawa Timur.	Negatif
Barokah Sedulur Ngaji, semoga Cak Imin sukses jadi wapres Jokowi dan Mbak Khofifah jadi menteri lagi.	Negatif
Khofifah Indar Parawansa-Emil Dardak dan Saifullah Yusuf (Gus Ipul)-Puti Guntur Soekarno beradu gagasan dan program.	Netral

C. Twitter

Twitter adalah situs jejaring sosial yang membatasi penggunanya untuk mengirim sebuah *tweet* (cuitan) dengan batas 280 karakter. Twitter mempunyai beberapa kesamaan dengan situs jejaring sosial lainnya seperti facebook yaitu berguna untuk menghubungkan antara pengguna satu dengan pengguna lainnya dan tidak terbatas oleh jarak. Twitter didirikan oleh Jack Dorsey dan diresmikan pada bulan Maret 2006. Twitter terus mengalami perkembangan jumlah pengguna yang pesat. Hal ini dibuktikan dengan jumlah penggunanya yang mencapai 500 juta pengguna pada Januari 2013 (Yulian, 2018). Kelebihan twitter dibandingkan

facebook adalah fitur *follow* yang memudahkan seseorang untuk terhubung dengan orang lain dengan jumlah yang tidak terbatas. Hal ini tentu lebih baik daripada facebook yang membatasi pertemanan hanya sejumlah 5000 orang.

Dalam twitter, ada beberapa istilah dasar yang harus dipahami agar seseorang dapat menggunakan situs jejaring sosial twitter (Maulana, 2012) diantaranya:

1. *Tweet*

   Tweet adalah apa yang seseorang tulis atau posting di dalam twitter (layaknya status facebook).

2. *Mention* (@)

   Mention adalah suatu cara untuk membuat link terhadap suatu akun twitter. Cara ini biasanya digunakan ketika seseorang membalas *tweet* atau ingin menandai suatu *tweet* kepada seseorang. Ketika seseorang membuat *tweet* yang memuat mention akun twitter orang lain, maka *tweet* tersebut akan muncul di beranda orang lain tersebut.

3. *Hashtag* (#)

   Cara menuliskan *hashtag* yaitu dengan cara menambahkan tanda # di depan topik yang ditulis. Penulisan topik di dalam hashtag tidak diperkenankan menggunakan spasi, jadi ketika suatu topik terdiri dari dua kata maka penulisannya digabung menjadi satu. *Hashtag* berfungsi agar pengguna lain yang melihat bisa mencari topik sejenis yang ditulis orang lain dengan mudah. Intinya untuk mengelompokkan suatu topik agar mudah dicari. *Hashtag* ini berhubungan dengan *trending topic*. Ketika suatu *hashtag* merupakan *hashtag* yang paling banyak diposting di

twitter, maka *hashtag* tersebut menjadi *trending topic*. Contoh: #pilkadajatim, #pemiluserentak

4. *ReTweet* (RT)

   ReTweet dalam bahasa Indonesia adalah mengulang suatu *tweet* seseorang maka *retweet* memiliki makna bahwa seseorang mengulang *tweet* orang lain untuk disampaikan kembali kepada *follower* nya. Dalam penelitian ini, data yang digunakan adalah data dalam twitter. Proses pencarian data dalam penelitian ini menggunakan metode *crawling*. *Crawling* itu prinsipnya seperti *search engine* tetapi mampu mencari data dalam jumlah yang besar dan menyimpannya.

D. Pemilihan Kepala Daerah 2018

Tahun 2018 ini merupakan tahun diadakannya pemilihan kepala daerah secara serentak hampir di seluruh daerah di Indonesia. Pemilihan kepala daerah adalah pemilihan pemimpin baru daerah secara langsung oleh masyarakat daerah tersebut. Sedangkan menurut PP Nomor 6 tahun 2005, pemilihan kepala daerah adalah pemilihan kepala daerah dan wakil kepala daerah sebagai sarana pelaksanaan kedaulatan rakyat di wilayah provinsi atau kabupaten ataupun kota berdasarkan Pancasila dan Undang-Undang Dasar Negara Republik Indonesia tahun 1945 untuk memilih kepala daerah dan wakil kepala daerah.

Joko J. Prihatantoro menyatakan bahwa pemilihan kepala daerah adalah rekrutmen politik yaitu penyeleksian rakyat terhadap tokoh-tokoh yang mencalonkan diri sebagai kepala daerah, baik gubernur/wakil gubernur maupun bupati/wakil bupati atau walikota/wakil walikota. Dalam kehidupan politik di

daerah, pemilihan kepala daerah merupakan salah satu kegiatan yang nilainya ekuivalen dengan pemilihan anggota DPRD. Ekuivalen tersebut ditunjukkan dengan kedudukan yang sejajar antara kepala daerah dan DPRD.

E. Klasifikasi

Klasifikasi merupakan kata serapan dari bahasa Inggris, *classification*. Menurut Kamus Besar Bahasa Indonesia, klasifikasi adalah proses pengelompokan benda berdasarkan ciri-ciri persamaan dan perbedaan. Menurut Saputro, klasifikasi adalah suatu penilaian terhadap sebuah data untuk dimasukkan ke dalam kelas tertentu dari beberapa kelas yang tersedia (Saputro, 2016).

Dalam klasifikasi, ada dua proses utama yang dilakukan. Pertama yaitu pembentukan model klasifikasi berdasarkan data-data yang sudah ada dan dijadikan sebagai memori. Kedua yaitu pengkategorian data baru yang didasarkan pada model klasifikasi yang telah dibentuk tersebut. Ada banyak metode yang bisa digunakan untuk melakukan klasifikasi. Metode yang sering digunakan yaitu *decision tree, artificial neural network, support vector machine*, dan *naive bayes classifier*.

F. Klasifikasi Dokumen Teks

Klasifikasi merupakan pembentukan suatu model yang akan digunakan untuk memprediksi suatu kategori (Juniawan, 2009). Klasifikasi memiliki dua tahap proses yaitu proses pembelajaran dan proses klasifikasi itu sendiri. Pada tahap pertama, algoritma klasifikasi membentuk suatu model klasifikasi dengan menganalisis data latih. Tahap ini disebut juga sebagai *supervised learning* karena setiap data latih telah memiliki label kelas masing-masing. Tahap ini bisa

diasumsikan sebagai pemetaan fungsi $y = f(x)$, yaitu menentukan label kelas y dari data x dengan fungsi pemetaan f. Tahap kedua yaitu penentuan kelas untuk data uji dengan menggunakan model klasifikasi yang telah terbentuk.

Klasifikasi dokumen teks tentu berbeda dengan klasifikasi data berupa numerik ataupun gambar. Dalam klasifikasi dokumen teks, yang menjadi variabel independennya adalah kata-kata yang berada pada dokumen teks tersebut. Hal ini menyebabkan klasifikasi dokumen teks mempunyai variabel independen yang sangat banyak, bergantung pada banyaknya kata yang ada pada kumpulan dokumen yang diklasifikasikan.

G. Teorema *Bayes*

Bayes adalah teknik prediksi berbasis probabilitas sederhana yang didasarkan pada teorema *Bayes* dengan asumsi independensi (tidak tergantung) yang kuat (naif). Secara sederhana, *naïve bayes* merupakan model yang menggunakan fitur independen sebagai hal yang utama. Maksud dari independen dalam *naïve bayes* yaitu bahwa fitur yang satu tidak ada kaitannya dengan ada atau tidaknya fitur yang lain pada data yang sama (Balagatabi, 2012).

Naïve bayes classifier didasarkan pada teorema bayes seperti Persamaan 2.1 dan 2.2 berikut ini:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

$$P(A \cap B) = P(A|B) \cdot P(B) \tag{2.2}$$

Keterangan:

- $P(A|B)$ adalah peluang terjadinya hipotesis A jika kejadian B sudah terjadi.

- $P(A \cap B)$ adalah peluang hipotesis A yang terjadi bersamaan dengan hipotesis B.
- $P(A)$ adalah peluang terjadinya hipotesis A tanpa melihat kondisi apapun.
- $P(B)$ adalah peluang terjadinya kejadian B tanpa melihat kondisi apapun.

Ide dasar dari teorema bayes adalah hasil dari peluang terjadinya hipotesis A dapat diperkirakan berdasarkan kejadian B yang diamati. Ada beberapa hal penting dari teorema bayes yang merupakan dasar dari *naïve bayes classifier* (Balagatabi, 2012) diantaranya:

1. Sebuah peluang awal atau $P(A)$ adalah peluang dari suatu hipotesis sebelum kejadian lain diamati.
2. Sebuah probabilitas akhir atau $P(A|B)$ adalah peluang dari suatu hipotesis setelah kejadian lain diamati.

H. *Naïve Bayes Classifier*

Metode *Naïve bayes* merupakan metode yang digunakan untuk memprediksi probabilitas (Rini, Farida, & Puspitasari, 2016). Metode ini memanfaatkan teori peluang yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Cara kerja metode ini yaitu dengan memprediksi peluang terjadinya kejadian di masa depan berdasarkan data yang ada sebelumnya. *Naïve bayes classifier* adalah konsep probabilitas yang bisa digunakan untuk penentuan kelompok kelas dokumen teks dan dapat mengolah data dalam jumlah besar dengan hasil akurasi yang tinggi (Lestari, Perdana, & Fauzi, 2017). Tingkat performa dari sistem klasifikasi yang dibuat menggunakan *naive bayes classifier* bergantung pada data yang dimiliki dan data yang dipilih sebagai data *training*. Jika data yang dipilih sebagai data *training*

bisa mewakili semua atau sebagian besar data yang dimiliki, maka sistem klasifikasi yang dibuat mempunyai performa yang bagus. Ketika sistem klasifikasi yang dibuat mempunyai performa yang bagus, maka sistem tersebut bisa digunakan untuk melakukan klasifikasi terhadap data yang lebih banyak.

Naïve bayes dituliskan dengan $P(X|Y)$ dimana X merupakan vektor masukan yang berisi fitur-fitur dan Y adalah label kelas. Notasi tersebut berarti probabilitas kelas Y diperoleh setelah fitur-fitur X diamati. Notasi ini biasa disebut probabilitas akhir (*posterior probability*) untuk Y, sedangkan $P(Y)$ merupakan probabilitas awal (*prior probability*) dari Y. Selama proses pelatihan data dilakukan pembelajaran probabilitas akhir $P(X|Y)$ pada model untuk setiap kombinasi dari X dan Y berdasarkan informasi dari data latih. Pelatihan data itu dilakukan untuk mendapatkan model klasifikasi yang nantinya digunakan untuk mengklasifikasikan data uji X dengan mencari nilai Y nya dengan memanfaatkan nilai $P(X|Y)$ yang didapat (Balagatabi, 2012).

Dalam mengklasifikasikan dokumen teks, ada beberapa proses yang harus dilakukan yaitu:

1. Mencari nilai peluang dari setiap kategori dokumen.
2. Mencari nilai peluang kemunculan dari masing-masing kata pada masing-masing kategori dokumen.
3. Menentukan kategori dokumen uji berdasarkan perhitungan dari proses pertama dan kedua.

I. Pre-proses Teks

Tahap pre-proses teks ini merupakan tahap yang dilakukan untuk mempersiapkan dokumen teks agar siap untuk diklasifikasikan. Tahap ini dilakukan karena dokumen teks yang ada biasanya tidak memiliki struktur yang pasti sehingga informasi di dalamnya tidak bisa diproses secara langsung. Alasan lain dilakukannya tahap ini adalah karena tidak semua kata dalam dokumen teks mencerminkan isi yang terkandung dalam dokumen tersebut. Tahap ini dilakukan sebelum dokumen teks diklasifikasikan dengan metode *naïve bayes classifier*. Tahap pre-proses teks ini meliputi *cleansing*, *case folding*, penghapusan *stopword*, tokenisasi, dan pembobotan kata.

1. *Cleansing*

Cleansing termasuk dalam pre-proses teks. *Cleansing* dilakukan untuk menghapus semua karakter html atau web yang tidak mempunyai makna dalam pengklasifikasian dokumen. Hal ini dikarenakan dalam sebuah *tweet* terkadang disertakan suatu alamat web yang jika tidak dihapus akan mengganggu proses klasifikasi.

2. *Case Folding*

Dalam melakukan klasifikasi dokumen, *case folding* merupakan salah satu hal yang harus dilakukan. Cara melakukan *case folding* yaitu dengan mengubah semua huruf menjadi huruf kecil. Hal ini dilakukan agar kata yang berada di awal kalimat mempunyai arti yang sama dengan kata di tengah atau akhir kalimat (Manning, Raghavan, & Schutze, 2009). Selain itu, di dalam

social media seperti twitter. Ada sebagian orang yang menulis *tweet* menggunakan huruf besar semua untuk melakukan penekanan makna. Sehingga sulit jika data tersebut diklasifikasikan nantinya. Setelah itu, dilakukan penghapusan tanda baca dan angka dan menggantinya dengan karakter spasi.

3. Penghapusan *Stopword*

Biasanya, kata-kata yang sering muncul di setiap kategori dokumen merupakan kata-kata yang tak bermakna dalam pengkategorian dokumen (Manning, Raghavan, & Schutze, 2009). Kata-kata tersebut biasanya disebut *stopword*. Kata-kata tersebut juga tak bisa dijadikan penciri suatu dokumen, sehingga kata-kata tersebut seharusnya dibuang. Kata-kata yang dibuang tersebut kemudian disimpan dalam daftar kata yang disebut *stoplist*. Contoh dari *stopword* adalah kata penghubung dan kata ganti.

4. Tokenisasi

Tokenisasi adalah proses untuk membagi dokumen teks yang dapat berupa kalimat atau paragraf menjadi token-token atau bagian-bagian tertentu (Manning, Raghavan, & Schutze, 2009). Tokenisasi ini dilakukan agar kata-kata yang ada pada kalimat atau paragraf bisa diberi bobot untuk setiap kata. Contohnya seperti kalimat “partai politik memainkan peran penting dalam proses pemilihan kepala daerah” setelah dilakukan tokenisasi maka menjadi “partai – politik – memainkan – peran – penting – dalam – proses – pemilihan – kepala - daerah” yang terdiri dari 10 token.

5. Pembobotan Kata

Hal yang perlu diperhatikan dalam pencarian informasi dari dokumen yang heterogen adalah pembobotan *term*. *Term* dapat berupa kata atau frase yang ada pada dokumen yang dapat digunakan untuk mengetahui konten dari dokumen tersebut. TF-IDF merupakan singkatan dari *Term Frequency – Inverse Document Frequency*. Jadi TF-IDF merupakan gabungan dari dua skema pembobotan yaitu TF atau *Term Frequency* dan IDF atau *Inverse Document Frequency* (Juniawan, 2009).

TF (*Term Frequency*) adalah frekuensi kemunculan *term* pada sebuah dokumen. Jadi, pembobotan TF didasarkan pada jumlah kemunculan *term* atau kata pada sebuah dokumen yang bersangkutan tersebut. Dalam TF, terdapat beberapa jenis formula yang digunakan untuk melakukan skoring antara lain:

1. TF biner, yang diperhatikan dalam TF biner hanya ada atau tidaknya suatu *term* pada dokumen yang bersangkuten, jika ada diberi nilai satu dan jika tidak ada diberi nilai 0.
2. TF murni, skor dari TF dinilai dari jumlah kemunculan suatu *term* pada dokumen. Contohnya, jika dalam sebuah dokumen kemunculan suatu *term* adalah 3 kali maka *term* tersebut akan bernilai 3 pada dokumen yang bersangkutan.
3. TF logaritmik, jenis TF ini digunakan untuk menghindari dominansi dokumen yang mengandung sedikit *term* dalam query, namun mempunyai frekuensi yang tinggi.

$$TF = \begin{cases} 1 + \log(\mathrm{f}_{t,\ d}) & ,(\mathrm{f}_{t,\ d}) > 0 \\ 0 & ,(\mathrm{f}_{t,\ d}) = 0 \end{cases} \qquad (2.3)$$

dimana $(f_{t,\ d})$ *adalah* frekuensi *term* pada dokumen.

4. TF normalisasi, skor dilakukan dengan cara membandingkan frekuensi sebuah *term* nilai maksimum dari keseluruhan atau kumpulan frekuensi *term* yang ada pada suatu dokumen.

IDF (*Inverse Document Frequency*) adalah sebuah pembobotan dari bagaimana *term* itu didistribusikan secara luas pada koleksi dokumen yang bersangkutan (Juniawan, 2009). Skor sebuah *term* pada IDF merupakan frekuensi kemunculan *term* tersebut di dalam semua dokumen. Semakin sedikit jumlah dokumen yang mengandung *term* yang dimaksud, maka semakin besar nilai IDFnya (Manning, Raghavan, & Schutze, 2009). Berikut ini adalah Persamaan 2.4 yang digunakan untuk menghitung nilai IDF:

$$\text{IDF}_i = \log_2(D/Df_j) \quad (2.4)$$

Dimana:

$IDF_i = bobot\ IDF\ term\ i$

$\text{D} = jumlah\ semua\ dokumen\ yang\ ada\ dalam\ database$

$DF_i = jumlah\ dokumen\ yang\ mengandung\ term\ i$

Contoh pembobotan IDF adalah sebagai berikut:

Misalkan sekumpulan dokumen opini dalam twitter berjumlah 1000 dokumen. Kemudian kata gubernur muncul pada 200 dokumen, maka pembobotan IDF dari kata gubernur adalah

$$\log_2\left(\frac{1000}{200}\right) = \log_2(5) = 2{,}32$$

Tabel 2.2 Contoh Pembobotan IDF

Term	DF	IDF
Gubernur	200	2,32
Politik	50	4,32
Rakyat	100	3,32

Tabel 2.2 menunjukkan contoh hasil perhitungan pembobotan kata dengan metode IDF. Metode perhitungan IDF ini selanjutnya dipadukan dengan pembobotan kata TF sehingga menjadi metode TF-IDF. Jenis TF yang biasa digunakan untuk pembobotan adalah TF murni. Namun, pada perhitungan pembobotan TF-IDF, dilakukan normalisasi TF murni. Normalisasi ini dilakukan dengan cara membagi nilai TF murni dengan jumlah *term* yang ada pada suatu dokumen. Oleh karena itu, rumus umum untuk pembobotan TF-IDF adalah penggabungan perhitungan normalisasi TF murni dengan formula IDF dengan cara mengalikan nilai normalisasi TF murni dan IDF nya. Berikut ini adalah Persamaan 2.5 yang digunakan untuk menghitung pembobotan TF-IDF dari suatu term:

$$W_{ij} = \frac{TF_{ij}}{K_j} \cdot IDF_i \tag{2.5}$$

$$W_{ij} = \frac{TF_{ij}}{K_j} \cdot \log_2\left(\frac{D}{DF_i}\right) \tag{2.6}$$

Dimana:

$W_{ij} = bobot\ term\ i\ terhadap\ dokumen\ j$

$K_j = jumlah\ semua\ term\ yang\ ada\ pada\ dokumen\ j$

$TF_{ij} = jumlah\ kemunculan\ term\ i\ pada\ dokumen\ j$

J. Evaluasi Sistem Klasifikasi

Sebuah sistem klasifikasi dibuat dengan harapan sistem tersebut dapat mengklasifikasikan semua data dengan benar (Prasetyo, 2014). Namun, tidak dapat dipungkiri bahwa sebuah sistem yang dibuat selalu terdapat kesalahan. Oleh karena itu, evaluasi sistem klasifikasi ini dilakukan untuk mengetahui seberapa baik sebuah metode itu digunakan untuk melakukan klasifikasi terhadap suatu data tertentu atau seberapa baik sistem klasifikasi yang telah dibuat.

Evaluasi merupakan hal yang sangat penting dalam setiap pembuatan sistem klasifikasi. Evaluasi sistem klasifikasi umumnya dilakukan dengan *confusion matrix*. *Confusion matrix* adalah sebuah matriks yang digunakan untuk membandingkan hasil klasifikasi dengan data yang asli (Abdillah, 2015). Konsep tentang *confusion matrix* ini diilustrasikan seperti Tabel 2.3:

Tabel 2.3 Tabel *Confusion Matrix*

Prediksi / Aktual	Positif	Negatif
Positif	True Positive	False Negative
Negatif	False Positive	True Negative

Pada *confusion matrix* terdapat empat jenis data yaitu *true positive, true negative, false positive,* dan *false negative* (Perdana, 2017). *True positive* adalah jumlah data yang mempunyai kelas positif yang diklasifikasikan sebagai kelas positif. Sedangkan *true negative* adalah jumlah data yang mempunyai kelas negatif yang diklasifikasikan sebagai kelas negatif. *True positive* dan *true negative* merupakan hasil klasifikasi yang benar. *False positive* adalah data yang sebenarnya merupakan kelas negatif tetapi diklasifikasikan sebagai kelas positif. Sedangkan

false negative merupakan data yang sebenarnya merupakan kelas positif tetapi diklasifikasikan sebagai kelas negatif.

Ada empat jenis evaluasi yang bisa digunakan untuk melakukan evaluasi sistem klasifikasi yaitu akurasi, presisi, *recall*, dan *f-measure*. Perhitungan empat jenis evaluasi ini dilakukan terhadap data uji yang dimiliki. Akurasi merupakan perbandingan antara banyaknya data yang diklasifikasikan secara benar dibagi oleh jumlah semua data. Sedangkan presisi merupakan perbandingan antara banyaknya *true positive* dibagi oleh jumlah semua data yang diklasifikasikan positif. *Recall* merupakan perbandingan antara banyaknya *true positive* dibagi oleh jumlah semua data yang berkategori positif. Sedangkan, *f-measure* merupakan perhitungan performa sistem klasifikasi yang didapat dengan mengkombinasikan perhitungan presisi dan *recall*.

Evaluasi kinerja sistem klasifikasi yang paling sering digunakan adalah akurasi. Akurasi bisa digunakan untuk mengukur kinerja sistem klasifikasi jika data yang digunakan mempunyai perbandingan jumlah kategori setiap data yang seimbang (Prasetyo, 2014). Namun, terkadang nilai akurasi tidak mennggambarkan performa sebenarnya dari sebuah sistem klasifikasi. Hal ini bisa terjadi jika perbandingan jumlah kategori setiap data yang ada sangat tidak seimbang (Prasetyo, 2014). Sebagai contoh, jika terdapat 95 data kategori positif dan hanya terdapat 5 kategori negatif dalam sebuah dataset, sebuah sistem klasifikasi mungkin akan mengklasifikasikan semua dataset sebagai data kategori positif. Akurasi dari sistem tersebut memang sebesar 95%, tetapi secara lebih detail sistem klasifikasi tersebut mempunyai *recognition rate* (*recall*) kelas positif sebesar 100% sedangkan

recognition rate dari kelas negatif hanya 0%. Hal ini menunjukkan bahwa sistem klasifikasi yang dibuat masih belum bisa mengenali salah satu kelas yaitu kelas negatif. Oleh karena itu, perhitungan akurasi saja tidak cukup untuk menggambarkan tingkat performa dari sistem klasifikasi yang memiliki proporsi kategori data yang tidak seimbang. Sehingga, dibutuhkan perhitungan evaluasi lain yang bisa menggambarkan performa sistem untuk masing-masing kategori yaitu presisi, *recall*, dan *f-measure*.

Untuk menghitung akurasi, presisi, dan *recall* dari sebuah sistem klasifikasi digunakan formula sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.7}$$

$$Presisi = \frac{TP}{TP + FP} \tag{2.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.9}$$

$$F - Measure = 2 \cdot \frac{Presisi \cdot Recall}{Presisi + Recall} \tag{2.10}$$

Dimana:

$TP = true\ positive$

$TN = true\ negative$

$FP = false\ positive$

$FN = false\ negative$

BAB III

METODE PENELITIAN

Pada bab ini, dijelaskan tentang jenis penelitian, data yang digunakan dalam penelitian, dan bagaimana proses yang dilakukan untuk menyelesaikan permasalahan dalam penelitin ini.

A. Jenis Penelitian

Pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif deskriptif. Pendekatan kuantitatif digunakan karena dalam penelitian ini, analisis data yang digunakan menggunakan pendekatan teori peluang *naive bayes*. Dalam teori ini, digunakan perhitungan numerik dari peluang suatu kejadian. Kejadian yang dimaksud dalam penelitian ini yaitu suatu opini termasuk dalam kategori tertentu. Selanjutnya, interpretasi hasil penelitian dilakukan dalam bentuk deskripsi. Oleh karenanya, jenis penelitian ini adalah penelitian kuantitatif yang bersifat deskriptif.

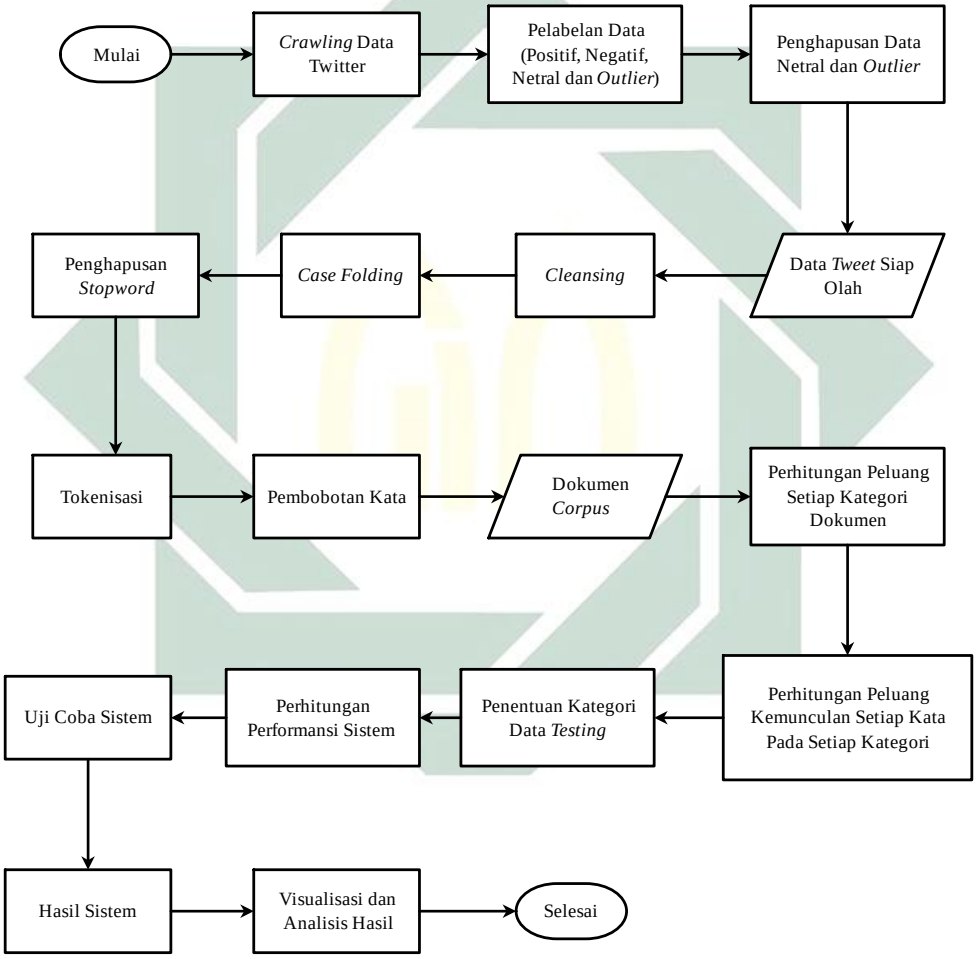
B. Jenis dan Sumber Data

Jenis data yang digunakan dalam penelitian ini adalah data sekunder yang berupa opini masyarakat di twitter tentang tokoh calon gubernur Jawa Timur periode 2018-2022 yang didapat dari situs www.api.twitter.com pada bulan April dan Mei 2018. Untuk mendapatkan data tersebut, dilakukan proses web *crawling* dari situs www.api.twitter.com dengan menggunakan aplikasi Rstudio. Secara sederhana, pada langkah ini dilakukan *crawling* dengan cara menentukan kata kunci apa yang ingin dicari atau bisa juga dengan menentukan akun twitter siapa yang

ingin dicari. Sebagai contoh, jika data yang dicari adalah opini tentang calon gubernur nomor urut satu, maka bisa dilakukan dengan menggunakan kata kunci "khofifah". Jika data yang dicari adalah opini tentang calon gubernur nomor urut dua, maka bisa dilakukan *crawling* dengan kata kunci "gus ipul".

C. Teknis Analisis Data

Dalam tahap ini dilakukan analisis terhadap data sekunder yang dimiliki.



Gambar 3.1 Teknis Analisis Data

Gambar 3.1 menjelaskan tentang teknis analisis terhadap data yang dimiliki. Dalam penelitian ini, terdapat tujuh proses yang dilakukan untuk menganalisis data dan mendapatkan hasil klasifikasi sentimen. Proses yang pertama yaitu proses

pengumpulan data dengan cara *crawling* yang digunakan untuk mendapatkan dataset berupa opini pada twitter. Proses yang kedua yaitu pelabelan data menjadi positif, negatif, netral, dan *outlier*. Proses yang ketiga yaitu penghapusan data netral dan *outlier*. Proses yang keempat yaitu pre-proses teks yang digunakan untuk mengubah dataset mentah menjadi data yang siap untuk diklasifikasikan. Proses yang kelima yaitu klasifikasi dengan menggunakan metode *naive bayes classifier*. Proses yang keenam yaitu uji coba sistem sehingga didapatkan hasil klasifikasi sentimen. Proses yang terakhir yaitu visualisasi dan analisis hasil sentimen masyarakat tentang pemilihan Kepala Daerah Jawa Timur pada tahun 2018. Agar lebih jelas, proses-proses tersebut dijelaskan seperti berikut ini:

1. *Crawling* Data Twitter

   Untuk mendapatkan data yang dibutuhkan dalam penelitian ini, dilakukan *crawling* data twitter. Langkah pertama yang dilakukan untuk melakukan *crawling* data twitter yaitu memiliki akun twitter. Dengan akun twitter yang sudah dimiliki tadi, langkah selanjutnya yaitu registrasi di https://dev.twitter.com/apps/new untuk mendapatkan kode akses API twitter. Setelah mendapatkan kode akses API twitter, selanjutnya digunakan *software* Rstudio untuk melakukan *crawling* data twitter. *Crawling* ini dilakukan dengan menentukan kata kunci apa yang dicari dan menentukan jumlah *tweet* yang diinginkan. Setelah itu, data hasil *crawling* disimpan dalam bentuk dokumen microsoft excel berekstensi .csv.

2. Pelabelan Data

Dari data yang sudah disimpan dalam bentuk dokumen *microsoft excel* tadi, dilakukan pelabelan data secara manual ke dalam empat kategori yaitu positif, negatif, netral, dan *outlier*. Pelabelan data ini dilakukan agar data siap diklasifikasikan.

3. Penghapusan Data Netral dan *Outlier*

Setelah dilakukan pelabelan data, dilakukan penghapusan data netral dan outlier atau yang tidak berhubungan dengan penelitian ini. Penghapusan data netral ini dilakukan karena keberadaan data netral yang sangat sedikit, sehingga keberadaannya akan mengganggu klasifikasi. Selain itu, data yang bisa digunakan sebagai peta kekuatan calon gubernur hanya data berlabel positif dan negatif saja. Salah satu contoh dari data outlier ini adalah data yang berbahasa Inggris.

4. *Cleansing*

Cleansing termasuk dalam pre-proses teks. *Cleansing* dilakukan untuk menghapus semua karakter html atau web yang tidak mempunyai makna dalam pengklasifikasian dokumen. Hal ini dikarenakan dalam sebuah *tweet* terkadang disertakan suatu alamat web yang jika tidak dihapus akan mengganggu proses klasikasi.

5. *Case Folding*

Case folding ini juga termasuk dalam pre-proses teks. *Case folding* adalah proses penghapusan tanda baca, angka, dan merubah huruf menjadi huruf kecil semua. Hal ini dilakukan agar dalam klasifikasi kata

yang sama yang penulisan huruf kapitalnya berbeda dianggap menjadi satu kata yang sama.

6. Penghapusan *Stopword*

   Stopword adalah kata-kata yang tidak mempunyai arti penting dalam pengkategorian dokumen. Hal ini dikarenakan *stopword* cenderung ada pada semua kategori dokumen. Contoh dari *stopword* adalah kata ganti dan kata penghubung.

7. Tokenisasi

   Tokenisasi adalah proses untuk membagi dokumen teks yang dapat berupa kalimat atau paragraf menjadi token-token atau bagian-bagian tertentu. Contohnya seperti kalimat "partai politik memainkan peran penting dalam proses pemilihan kepala daerah" setelah dilakukan tokenisasi maka menjadi "partai-politik-memainkan-peran-penting-dalam-proses-pemilihan-kepala-daerah" yang terdiri dari 10 token.

8. Pembobotan Kata

   Setelah dilakukan tokenisasi, setiap kata dalam dokumen diberi bobot dengan dua metode yang berbeda, pertama yaitu TF murni dan yang kedua yaitu TF-IDF seperti yang sudah dijelaskan di tinjauan pustaka. Dalam proses ini, semua kata diproses sehingga setiap kata yang ada pada dokumen tersebut mempunyai bobot sendiri-sendiri. Setelah proses ini selesai, maka kumpulan dokumen (dokumen *corpus*) telah siap untuk dilakukan *training* untuk proses klasifikasi.

9. Klasifikasi dengan *Naive Bayes*

Proses selanjutnya yaitu klasifikasi itu sendiri. Dalam proses ini, dilakukan pelatihan terhadap data teks yang sudah dilakukan pre-proses dan seleksi fitur. Pelatihan itu dilakukan untuk membentuk model klasifikasi yang nantinya akan digunakan untuk mengklasifikasikan data uji X agar mendapatkan label Y.

Secara umum proses yang dilakukan dalam klasifikasi naive bayes adalah sebagai berikut:

1. Mencari nilai peluang dari setiap kategori dokumen.
2. Mencari nilai peluang kemunculan dari masing-masing kata pada masing-masing kategori dokumen.
3. Menentukan kategori dokumen uji berdasarkan perhitungan dari proses pertama dan kedua.

10. Perhitungan Performa Sistem

Setelah data uji selesai diklasifikasikan, maka dilakukan perhitungan performa sistem yang terdiri dari akurasi, presisi, *recall*, dan *f-measure* dengan *confusion matrix*.

11. Uji Coba Sistem

Proses selanjutnya yaitu uji coba sistem. Dalam proses ini, sistem yang sudah dibuat pada proses keempat diaplikasikan untuk data dengan skala yang lebih besar pada dokumen twitter. Setelah itu, didapatkan hasil klasifikasi berupa sentimen masyarakat terhadap calon gubernur Jawa

Timur 2018 yang nantinya bisa digunakan untuk evaluasi kampanye masing-masing pasangan calon.

12. Visualisasi dan Analisis Hasil

    Setelah didapatkan hasil klasifikasi sentimen masyarakat terhadap calon gubernur Jawa Timur. Langkah selanjutnya yaitu dilakukan visualisasi berupa pembuatan *wordcloud*. Kegunaan dari dibuatnya *wordcloud* ini adalah untuk mengetahui kata-kata mana yang sering muncul yang berpengaruh besar dalam pembuatan model klasifikasi sistem. Selain itu, data yang didapat juga divisualisasikan ke dalam bentuk grafik agar lebih mudah dipahami.

BAB IV

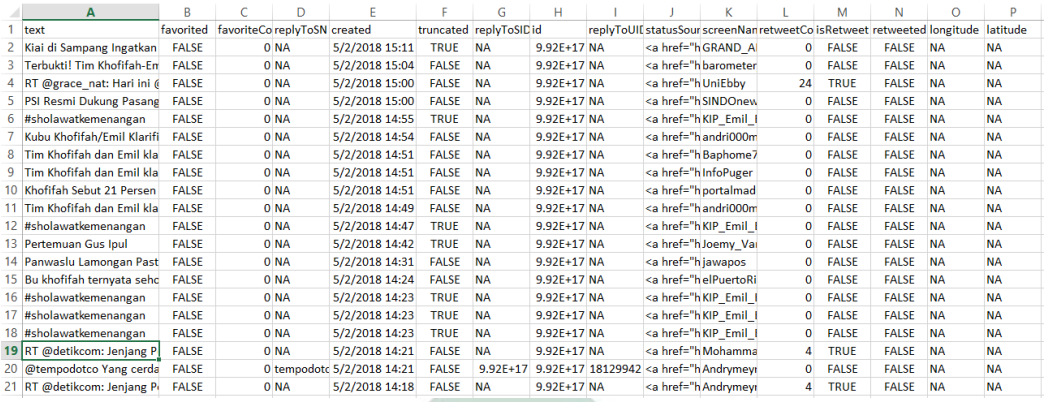
HASIL DAN PEMBAHASAN

Pada bab ini, dijelaskan tentang hasil yang didapatkan dari penelitian ini dari proses yang pertama sampai proses yang terakhir.

A. *Crawling* Data Twitter

Langkah pertama yang dilakukan pada penelitian ini yaitu *crawling* data twitter. *Crawling* data twitter ini harus dilakukan secara kontinu. Hal ini dikarenakan data yang bisa dicari pada proses *crawling* terbatas pada sepuluh hari terakhir dari waktu pencarian. Pada tahap ini, pencarian data twitter dibagi menjadi dua bagian. Bagian pertama yaitu pencarian data yang berhubungan dengan Khofifah Indar Parawansa, calon gubernur dengan nomor urut satu. Bagian kedua yaitu pencarian data yang berhubungan dengan Saifullah Yusuf, calon gubernur dengan nomor urut dua.

Crawling data twitter ini dilakukan pada tanggal 1 Mei sampai 7 Juni 2018. Untuk mencari data yang berhubungan dengan calon gubernur dengan nomor urut satu, Khofifah Indar Parawansa, digunakan kata kunci “khofifah”, sehingga nanti akan termuat semua *tweet* yang berhubungan dengan khofifah dalam sepuluh hari terakhir. Dari proses *crawling* untuk mencari data yang berhubungan dengan calon gubernur nomor urut satu, didapat data sebanyak 20937 data. Data tersebut merupakan data yang terdiri dari beberapa kategori diantaranya, positif, negatif, netral, dan outlier. Berikut ini adalah Gambar 4.1 yang merupakan contoh hasil *crawling* data twitter yang berhubungan dengan Khofifah Indar Parawansa yang tersimpan dalam file microsoft excel berekstensi .csv (*Comma Delimited*):



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	text	favorited	favoriteCo	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSour	screenNan	retweetCo	isRetweet	retweeted	longitude	latitude
2	Kiai di Sampang Ingatkan	FALSE	0	NA	5/2/2018 15:11	TRUE	NA	9.92E+17	NA	<a href="h	GRAND_A	0	FALSE	FALSE	NA	NA
3	Terbukti! Tim Khofifah-En	FALSE	0	NA	5/2/2018 15:04	FALSE	NA	9.92E+17	NA	<a href="h	barometer	0	FALSE	FALSE	NA	NA
4	RT @grace_nat: Hari ini @	FALSE	0	NA	5/2/2018 15:00	FALSE	NA	9.92E+17	NA	<a href="h	UniEbby	24	TRUE	FALSE	NA	NA
5	PSI Resmi Dukung Pasang	FALSE	0	NA	5/2/2018 15:00	FALSE	NA	9.92E+17	NA	<a href="h	SINDOnew	0	FALSE	FALSE	NA	NA
6	#sholawatkemenangan	FALSE	0	NA	5/2/2018 14:55	TRUE	NA	9.92E+17	NA	<a href="h	KIP_Emil_I	0	FALSE	FALSE	NA	NA
7	Kubu Khofifah/Emil Klarifi	FALSE	0	NA	5/2/2018 14:54	FALSE	NA	9.92E+17	NA	<a href="h	andri000m	0	FALSE	FALSE	NA	NA
8	Tim Khofifah dan Emil kla	FALSE	0	NA	5/2/2018 14:51	FALSE	NA	9.92E+17	NA	<a href="h	Baphome7	0	FALSE	FALSE	NA	NA
9	Tim Khofifah dan Emil kla	FALSE	0	NA	5/2/2018 14:51	FALSE	NA	9.92E+17	NA	<a href="h	InfoPuger	0	FALSE	FALSE	NA	NA
10	Khofifah Sebut 21 Persen	FALSE	0	NA	5/2/2018 14:51	FALSE	NA	9.92E+17	NA	<a href="h	portalmad	0	FALSE	FALSE	NA	NA
11	Tim Khofifah dan Emil kla	FALSE	0	NA	5/2/2018 14:49	FALSE	NA	9.92E+17	NA	<a href="h	andri000m	0	FALSE	FALSE	NA	NA
12	#sholawatkemenangan	FALSE	0	NA	5/2/2018 14:47	TRUE	NA	9.92E+17	NA	<a href="h	KIP_Emil_I	0	FALSE	FALSE	NA	NA
13	Pertemuan Gus Ipul	FALSE	0	NA	5/2/2018 14:42	TRUE	NA	9.92E+17	NA	<a href="h	Joemy_Va	0	FALSE	FALSE	NA	NA
14	Panwaslu Lamongan Past	FALSE	0	NA	5/2/2018 14:31	FALSE	NA	9.92E+17	NA	<a href="h	jawapos	0	FALSE	FALSE	NA	NA
15	Bu khofifah ternyata sehc	FALSE	0	NA	5/2/2018 14:24	FALSE	NA	9.92E+17	NA	<a href="h	elPuertoRi	0	FALSE	FALSE	NA	NA
16	#sholawatkemenangan	FALSE	0	NA	5/2/2018 14:23	TRUE	NA	9.92E+17	NA	<a href="h	KIP_Emil_I	0	FALSE	FALSE	NA	NA
17	#sholawatkemenangan	FALSE	0	NA	5/2/2018 14:23	TRUE	NA	9.92E+17	NA	<a href="h	KIP_Emil_I	0	FALSE	FALSE	NA	NA
18	#sholawatkemenangan	FALSE	0	NA	5/2/2018 14:23	TRUE	NA	9.92E+17	NA	<a href="h	KIP_Emil_I	0	FALSE	FALSE	NA	NA
19	RT @detikcom: Jenjang P	FALSE	0	NA	5/2/2018 14:21	FALSE	NA	9.92E+17	NA	<a href="h	Mohamma	4	TRUE	FALSE	NA	NA
20	@tempodotco Yang cerda	FALSE	0	tempodotc	5/2/2018 14:21	FALSE	9.92E+17	9.92E+17	18129942	<a href="h	Andrymeyr	0	FALSE	FALSE	NA	NA
21	RT @detikcom: Jenjang P	FALSE	0	NA	5/2/2018 14:18	FALSE	NA	9.92E+17	NA	<a href="h	Andrymeyr	4	TRUE	FALSE	NA	NA

Gambar 4.1 *Sample* Data untuk Khofifah Indar Parawansa

Pada Gambar 4.1, terlihat bahwa dari hasil *crawling* terdapat enam belas variabel data yang didapat. Data yang digunakan untuk proses selanjutnya yaitu data *text*, *screenName*, dan *retweetCount*. Data *text* digunakan untuk klasifikasi sentimen sedangkan data *screenName* dan *retweetCount* ditampilkan dalam grafik.

Selanjutnya, untuk mencari data yang berhubungan dengan calon gubernur nomor urut dua, Saifullah Yusuf, digunakan kata kunci "gus ipul", sehingga nanti termuat semua data yang berhubungan dengan Saifullah Yusuf selama sepuluh hari terakhir. Dari proses *crawling* data ini didapat data sebanyak 16098 data. Berikut ini Gambar 4.2 yang merupakan contoh data hasil *crawling* untuk Saifullah Yusuf:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	text	favorited	favoriteCo	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSou	screenNan	retweetCo	isRetweet	retweeted	longitude	latitude
2	RT @KontanNews:	FALSE	0	NA	5/2/2018 15:27	FALSE	NA	9.92E+17	NA	<a href="h	Leecque	1	TRUE	FALSE	NA	NA
3	Masalah lapangan	FALSE	0	NA	5/2/2018 15:17	TRUE	NA	9.92E+17	NA	<a href="h	ninaherlir	0	FALSE	FALSE	NA	NA
4	Gus Ipul bertemu F	FALSE	0	NA	5/2/2018 15:17	TRUE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
5	Semangat terus	FALSE	0	NA	5/2/2018 15:16	TRUE	NA	9.92E+17	NA	<a href="h	sabiahusn	0	FALSE	FALSE	NA	NA
6	Kemakmuran	FALSE	0	NA	5/2/2018 15:16	TRUE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
7	Saya doakan yang	FALSE	0	NA	5/2/2018 15:15	FALSE	NA	9.92E+17	NA	<a href="h	pratamafa	0	FALSE	FALSE	NA	NA
8	Lanjutkan Gus Ipul.	FALSE	0	NA	5/2/2018 15:15	TRUE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
9	Siap dukung gus	FALSE	0	NA	5/2/2018 15:14	TRUE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
10	PDI-P Surabaya me	FALSE	3	NA	5/2/2018 15:10	FALSE	NA	9.92E+17	NA	<a href="h	KontanNe	1	FALSE	FALSE	NA	NA
11	Selamat Hari	FALSE	0	NA	5/2/2018 15:08	FALSE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
12	pemimpin yang me	FALSE	0	NA	5/2/2018 15:08	FALSE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
13	Pendidikan kuwi p	FALSE	0	NA	5/2/2018 15:08	FALSE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
14	Bangsa yang maju a	FALSE	0	NA	5/2/2018 15:07	TRUE	NA	9.92E+17	NA	<a href="h	NormaEka	0	FALSE	FALSE	NA	NA
15	RT @beritamalangl	FALSE	0	NA	5/2/2018 15:06	FALSE	NA	9.92E+17	NA	<a href="h	BanggaBai	5	TRUE	FALSE	NA	NA
16	RT @beritamalangl	FALSE	0	NA	5/2/2018 15:05	FALSE	NA	9.92E+17	NA	<a href="h	ramdani_a	5	TRUE	FALSE	NA	NA
17	RT @beritamalangl	FALSE	0	NA	5/2/2018 15:04	FALSE	NA	9.92E+17	NA	<a href="h	BanggaBai	5	TRUE	FALSE	NA	NA
18	RT @beritamalangl	FALSE	0	NA	5/2/2018 15:04	FALSE	NA	9.92E+17	NA	<a href="h	AzkaYusuf	5	TRUE	FALSE	NA	NA
19	(8/10) penerapan k	FALSE	0	NA	5/2/2018 15:03	TRUE	NA	9.92E+17	NA	<a href="h	Abyantorc	0	FALSE	FALSE	NA	NA
20	RT @beritamalangl	FALSE	0	NA	5/2/2018 15:03	FALSE	NA	9.92E+17	NA	<a href="h	ramdani_a	5	TRUE	FALSE	NA	NA
21	RT @beritamalangl	FALSE	0	NA	5/2/2018 15:02	FALSE	NA	9.92E+17	NA	<a href="h	KhairaniA	5	TRUE	FALSE	NA	NA

Gambar 4.2 *Sample* Data untuk Saifullah Yusuf

Pada Gambar 4.2, seperti sebelumnya, data yang digunakan untuk proses selanjutnya yaitu data *text*, *screenName*, dan *retweetCount*.

B. Seleksi dan Pelabelan Data

Langkah kedua dalam penelitian ini adalah seleksi dan pelabelan data. Pada tahap ini, data yang didapat dari proses *crawling* twitter diseleksi secara manual sehingga data yang diberi label merupakan data yang mewakili semua data yang ada. Selanjutnya, data yang telah diseleksi dalam proses sebelumnya diberi label untuk setiap datanya. Data yang diperoleh diberi empat jenis label yang berbeda yaitu positif, negatif, netral, dan outlier. Data yang diberi label ini tentu bukan semua data yang diperoleh dari proses *crawling*. Data yang diberi label ini adalah data yang digunakan untuk pembuatan sistem. Jadi, dari puluhan ribu data yang didapat dari proses *crawling* dipilih ribuan data yang mewakili semua data yang ada tersebut. Hal ini dilakukan karena pemberian label untuk semua data membutuhkan waktu yang sangat lama. Selain itu, sistem yang dibuat dengan data *training* yang sangat banyak menjadi berat untuk dijalankan.

Tabel 4.1 berisi contoh pemberian label atau kategori secara manual.

Tabel 4.1 Contoh Pemberian Label pada Data

No	Data	Label Kelas
1.	Kaum buruh di sidoarjo dukung khofifah – emil dardak	Positif
2.	Ingin menangkan pilgub, tim khofifah diduga pasang APK di musholla	Negatif
3.	Peduli kaum perempuan, gus ipul dapat aisyiyah award 2018	Positif
4.	Nahdliyin ragu dukung gus ipul karena terkesan partai merah yang dominan	Negatif

C. Penghapusan Data Outlier dan Netral

Setelah data selesai diseleksi dan diberi label untuk masing-masing data. Langkah selanjutnya yaitu penghapusan data yang tidak dibutuhkan secara manual dalam proses pembentukan sistem klasifikasi. Alasan penghapusan data outlier adalah karena data outlier merupakan data yang tidak berhubungan dengan data yang dibutuhkan dalam penelitian ini dan apabila dibiarkan hanya akan mengganggu sistem untuk mengenali label kelas data lain.

Sedangkan alasan penghapusan data netral adalah karena jumlah data netral yang sangat sedikit jumlahnya sehingga ketika dimasukkan dalam sistem nanti dikhawatirkan akan mengganggu proses pengenalan label kelas data positif dan negatif. Hal lain yang melatarbelakangi penghapusan data netral adalah karena data netral merupakan data yang tidak tidak dibutuhkan dalam analisis kekuatan calon gubernur.

Tabel 4.2 berisi contoh data outlier dan data netral yang dihapus dari data yang digunakan sebagai data *training*.

Tabel 4.2 Contoh Data Outlier dan Netral

No	Data	Label Kelas
1.	Elektabilitas Khofifah-Emil dan Gus Ipul-Puti seimbang.	Netral
2.	Siapapun yang menjadi gubernur, entah bu khofifah atau gus ipul, semoga bisa membawa jatim menjadi lebih baik.	Netral
3.	@khofifah_jule Maacih	Outlier
4.	https://t.co/sk116uwjbl	Outlier

D. Pre-proses Teks

Setelah dilakukan penghapusan data outlier dan data netral, langkah selanjutnya yaitu pre-proses teks. Pre-proses teks ini terdiri dari beberapa tahapan. Tahapan dalam pre-proses teks ini yaitu *cleansing*, *case folding*, penghapusan *stopword*.

Tahap pertama dalam pre-proses ini adalah *cleansing*. Dalam *cleansing* ini, semua karakter web atau html dihapus. Hal ini dikarenakan karakter html tersebut tidak mempunyai makna dalam proses pengklasifikasian dokumen. Tabel 4.3 memberikan gambaran perbedaan data sebelum dan sesudah dilakukan *cleansing*.

Tabel 4.3 Contoh Penghapusan Karakter Web atau Html

No	Sebelum Data Dilakukan *Cleansing*	Setelah Data Dilakukan *Cleansing*
1	@PBB2019: Yusril Ihza Mahendra Pimpin Deklarasi Partai Bulan Bintang Dukung Khofifah – Emil https://t.co/CHBseKxNJY	@PBB2019: Yusril Ihza Mahendra Pimpin Deklarasi Partai Bulan Bintang Dukung Khofifah – Emil
2	Diusung PPP, Khofifah Raih Dukungan RKH M Syamsul Arifin Pamekasan https://t.co/CXMgIKUpT1	Diusung PPP, Khofifah Raih Dukungan RKH M Syamsul Arifin Pamekasan
3	@liputan6dotcom: Dukung Gus Ipul-Puti, Anak Milenial Sampang Buat Lagu https://t.co/ydSe22cqwd	@liputan6dotcom: Dukung Gus Ipul-Puti, Anak Milenial Sampang Buat Lagu
4	Gus Ipul: Banyak PR yang Harus Digarap https://t.co/LdajX7zUOM	Gus Ipul: Banyak PR yang Harus Digarap

Tahap selanjutnya dalam pre-proses yaitu *case folding*. Pada tahap ini, semua huruf yang ada pada dokumen dirubah ke bentuk standar, yaitu *lower case*. Selain itu, semua karakter selain huruf yaitu tanda baca dan angka dihapus. Tabel 4.4 memberikan gambaran perbedaan data sebelum dan sesudah dilakukan *case folding*.

Tabel 4.4 Contoh *Case Folding*

No	Sebelum Data Dilakukan *Case Folding*	Setelah Data Dilakukan *Case Folding*
1	@PBB2019: Yusril Ihza Mahendra Pimpin Deklarasi Partai Bulan Bintang Dukung Khofifah – Emil	pbb yusril ihza mahendra pimpin deklarasi partai bulan bintang dukung khofifah emil
2	Diusung PPP, Khofifah Raih Dukungan RKH M Syamsul Arifin Pamekasan	diusung ppp khofifah raih dukungan rkh m syamsul arifin pamekasan
3	@liputan6dotcom: Dukung Gus Ipul-Puti, Anak Milenial Sampang Buat Lagu	liputandotcom dukung gus ipul puti anak milenial sampang buat lagu
4	Gus Ipul: Banyak PR yang Harus Digarap	gus ipul banyak pr yang harus digarap

Tahap selanjutnya yaitu penghapusan *stopword*. Pada tahap ini, semua kata yang masuk dalam daftar *stopword* bahasa Indonesia dihapus. Penghapusan *stopword* ini dilakukan karena kata yang masuk dalam kategori *stopword* tidak bisa digunakan sebagai ciri kategori dokumen tertentu. Tabel 4.5 menunjukkan perbedaan data yang belum dilakukan penghapusan *stopword* dan sudah dilakukan penghapusan *stopword*.

Tabel 4.5 Contoh Penghapusan *Stopword*

No	Sebelum Dilakukan Penghapusan *Stopword*	Setelah Dilakukan Penghapusan *Stopword*
1	gus ipul banyak pr yang harus digarap	gus ipul pr digarap
2	kami selalu mendoakan yang terbaik untuk gus ipul dan mbak puti	mendoakan terbaik gus ipul mbak puti
3	kita lihat bu khofifah dengan energi yang seakan tidak ada habisnya terus berkeliling kemarin dari sendang biru	lihat bu khofifah energi seakan tidak habisnya berkeliling kemarin sendang biru
4	relawan gus dur siap sumbang juta suara untuk khofifah	relawan gus dur siap sumbang juta suara khofifah

E. Tokenisasi dan Pembobotan Kata

Langkah selanjutnya yaitu tokenisasi dan pembobotan kata. Tokenisasi dilakukan untuk melakukan partisi terhadap dokumen yang berupa kalimat menjadi kata-kata. Hal ini dilakukan agar kata-kata yang ada pada dokumen twitter bisa diberikan bobot untuk masing-masing kata. Tabel 4.6 berisi contoh tokenisasi dari sebuah *tweet*.

Tabel 4.6 Tokenisasi Kata

No	Data	Setelah Dilakukan Proses Tokenisasi
1	gus ipul pr digarap	gus – ipul – pr – digarap
2	mendoakan terbaik gus ipul mbak puti	mendoakan – terbaik – gus – ipul – mbak – puti
3	lihat bu khofifah energi seakan tidak ada habisnya berkeliling kemarin sendangbiru	lihat – bu – khofifah – energi – seakan – tidak – habisnya – berkeliling – kemarin – sendangbiru
4	relawan gus dur siap sumbang juta suara khofifah	relawan – gus – dur – siap – sumbang – juta – suara – khofifah

Setelah dilakukan tokenisasi, langkah selanjutnya yaitu diberi bobot untuk tiap kata. Ada dua metode yang digunakan untuk melakukan pembobotan kata. Metode yang pertama yaitu metode TF. Jenis TF yang digunakan adalah TF murni. Metode pembobotan kata ini merupakan metode yang sederhana karena bobot dari setiap kata ditentukan oleh banyak kemunculan kata tersebut pada *tweet* yang bersangkutan.

Metode pembobotan kata yang kedua yaitu TF-IDF. Metode pembobotan kata ini merupakan pengembangan dari metode TF. Setelah dilakukan pembobotan TF, bobot dari setiap kata itu juga dikalikan dengan IDF atau banyak kemunculan kata

tersebut pada kumpulan dokumen yang bersangkutan. Semakin jarang sebuah kata muncul pada dokumen, maka nilai IDF nya semakin besar.

Dua metode pembobotan kata yang digunakan untuk pembuatan sistem mempunyai performa yang sedikit berbeda. Dalam penelitian ini, metode pembobotan TF sedikit lebih unggul dibandingkan dengan pembobotan kata TF-IDF. Untuk data calon gubernur nomor urut satu, pembobotan kata TF dan TF-IDF mempunyai performa yang sama. Hal ini terlihat dari nilai akurasi, presisi, *recall*, dan *f-measure* dari kedua pembobotan kata yang mempunyai nilai sama yaitu berturut-turut 98,99%; 97,78%; 93,44%; dan 95,61%. Tetapi, untuk data calon gubernur dengan nomor urut dua, pembobotan kata TF sedikit lebih unggul dari pembobotan kata TF-IDF. Untuk pembobotan kata TF didapatkan akurasi, presisi, *recall*, dan *f-measure* berturut-turut yaitu 98,95%; 98,55%; 97,78%; dan 98,17% . Sedangkan pembobotan kata TF-IDF mempunyai akurasi, presisi, *recall*, dan *f-measure* berturut-turut yaitu 98,25%; 97,32%; 96,57%; dan 96,95%.

F. Pembuatan Sistem Klasifikasi

Setelah semua kata yang ada pada data yang digunakan sebagai data *training* sudah diberi bobot dengan metode TF dan TF-IDF, maka langkah selanjutnya adalah pembuatan sistem klasifikasi menggunakan metode *naive bayes classifier*. Seperti yang sudah dijelaskan pada tinjauan pustaka, cara kerja *naive bayes classifier* ini didasarkan pada peluang kemunculan kata yang sudah diberi bobot tadi.

Pada pembuatan sistem klasifikasi, tahap pertama yang harus dilakukan adalah membagi data *training* dan *testing*. Pembagian data ini tidak bisa dilakukan secara acak karena akan menyebabkan ketidakseimbangan proporsi yang ada pada data *training* dan data *testing*. Selain itu, data *training* yang digunakan juga harus merupakan data yang mewakili sebagian besar data yang ada, sehingga tidak bisa dilakukan secara acak. Proporsi pembagian data yang sering digunakan yaitu 80% sebagai data *training* dan 20% sebagai data *testing*. Selain itu, ada juga pembagian dengan proporsi 75% data *training* dan 25% data *testing*. Pada sistem ini, digunakan pembagian data *training* 80% dan data *testing* 20%.

Pada sistem yang pertama, data yang digunakan berjumlah 2497 data yang dibagi menjadi data *training* dan data *testing*. Data *training* terdiri dari 2000 data yang mempunyai kategori positif berjumlah 1906 data dan kategori negatif berjumlah 94 data. Sedangkan data *testing* berjumlah 497 data yang terdiri dari 469 data berkategori positif dan 28 data berkategori negatif.

Pada sistem yang kedua, data yang digunakan berjumlah 1487 data yang dibagi menjadi data *training* dan data *testing*. Data *training* terdiri dari 1200 data yang mempunyai kategori positif berjumlah 1132 data dan kategori negatif berjumlah 68 data. Sedangkan data *testing* berjumlah 287 yang terdiri dari 238 data berkategori positif dan 49 data berkategori negatif.

Berikut ini adalah contoh perhitungan manual *naive bayes classifier* berdasarkan kata yang sudah diberi pembobotan TF murni. Contoh pertama yaitu *tweet* dengan isi"PKH Plus untuk lansia dan disabilitas, Sebut Khofifah sebagai

salah satu Program Handal Untuk Jawa Timur #WisWayahe" yang dalam proses *testing* masuk ke dalam kategori positif.

Untuk membuat perhitungan *naive bayes* secara manual, tahap pertama yang harus dilakukan yaitu menghitung frekuensi kemunculan term pada contoh pertama pada data *training*. Tabel 4.7 menunjukkan frekuensi kemunculan *term* yang ada pada contoh pertama pada kumpulan data *training*.

Tabel 4.7 Tabel Frekuensi Kemunculan *Term* untuk Contoh Pertama

No	*Term*	Frekuensi Kategori *Training*	
		Positif	Negatif
1.	Pkh	38	7
2.	plus	5	0
3.	lansia	14	0
4.	disabilitas	1	0
5.	sebut	30	1
6.	khofifah	1429	75
7.	salah	16	1
8.	Satu	41	1
9.	program	59	2
10.	handal	0	0
11.	Jawa	197	3
12.	timur	186	3
13.	wiswayahe	23	1

Selain perhitungan jumlah *term* yang ada pada pada Tabel 4.7, harus juga diketahui jumlah semua *term* yang ada pada dokumen positif dan dokumen negatif. Pada dokumen positif terdapat 3343 *term* dan pada dokumen negatif terdapat 396 *term*.

Langkah kedua yaitu menghitung peluang kemunculan masing-masing kelas. Data yang digunakan *training* berjumlah 2000 data dimana 1906 data berkategori positif dan 94 data berkategori negatif. Oleh karena itu, peluang kemunculan kategori positif adalah:

$$P(positif) = \frac{Jumlah\ data\ positif}{Jumlah\ semua\ data}$$

$$P(positif) = \frac{1906}{2000}$$

$$P(positif) = 9{,}53 \cdot 10^{-1}$$

Sedangkan untuk kategori kelas negatif, peluang kemunculannya adalah sebagai berikut:

$$P(negatif) = \frac{Jumlah\ data\ negatif}{Jumlah\ semua\ data}$$

$$P(negatif) = \frac{94}{2000}$$

$$P(negatif) = 4{,}70 \cdot 10^{-2}$$

Langkah ketiga yaitu menghitung peluang kemunculan *term* pada masing-masing kategori dokumen. Berikut ini adalah contoh perhitungan peluang setiap *term* pada setiap kategori dokumen:

1. *term* “pkh”

$$P(\text{“}pkh\text{"}|positif) = \frac{n_k + 1}{n_{positif}} = \frac{38 + 1}{3343} = \frac{39}{3343} = 1{,}17 \cdot 10^{-2}$$

$$P(\text{“}pkh\text{"}|negatif) = \frac{n_k + 1}{n_{negatif}} = \frac{7 + 1}{396} = \frac{8}{396} = 2{,}02 \cdot 10^{-2}$$

2. *term* “plus”

$$P(\text{“}plus\text{"}|positif) = \frac{n_k + 1}{n_{positif}} = \frac{5+1}{3343} = \frac{6}{3343} = 1{,}79 \cdot 10^{-3}$$

$$P(\text{“}plus\text{"}|negatif) = \frac{n_k + 1}{n_{negatif}} = \frac{0+1}{396} = \frac{1}{396} = 2{,}53 \cdot 10^{-3}$$

3. *term* “lansia”

$$P(\text{“}lansia\text{"}|positif) = \frac{n_k + 1}{n_{positif}} = \frac{14+1}{3343} = \frac{15}{3343} = 4{,}49 \cdot 10^{-3}$$

$$P(\text{“}lansia\text{"}|negatif) = \frac{n_k + 1}{n_{negatif}} = \frac{0+1}{396} = \frac{1}{396} = 2{,}53 \cdot 10^{-3}$$

Dan seterusnya.

Tabel 4.8 menunjukkan hasil lengkap perhitungan manual langkah ketiga.

Tabel 4.8 Hasil Perhitungan Manual untuk Contoh Pertama

No	*Term*	Frekuensi Kategori *Training*		Frekuensi Dokumen Uji	Peluang Kemunculan *Term*	
		Positif	Negatif		Positif	Negatif
1.	Pkh	38	7	1	$1{,}17 \cdot 10^{-2}$	$2{,}02 \cdot 10^{-2}$
2.	Plus	5	0	1	$1{,}79 \cdot 10^{-3}$	$2{,}53 \cdot 10^{-3}$
3.	Lansia	14	0	1	$4{,}49 \cdot 10^{-3}$	$2{,}53 \cdot 10^{-3}$
4.	disabilitas	1	0	1	$5{,}98 \cdot 10^{-4}$	$2{,}53 \cdot 10^{-3}$
5.	Sebut	30	1	1	$9{,}27 \cdot 10^{-3}$	$5{,}05 \cdot 10^{-3}$
6.	khofifah	1429	75	1	$4{,}28 \cdot 10^{-1}$	$1{,}92 \cdot 10^{-1}$
7.	Salah	16	1	1	$5{,}09 \cdot 10^{-3}$	$5{,}05 \cdot 10^{-3}$
8.	Satu	41	1	1	$1{,}26 \cdot 10^{-2}$	$5{,}05 \cdot 10^{-3}$
9.	program	59	2	1	$1{,}79 \cdot 10^{-2}$	$7{,}58 \cdot 10^{-3}$
10.	Handal	0	0	1	$2{,}99 \cdot 10^{-4}$	$2{,}53 \cdot 10^{-3}$
11.	Jawa	197	3	1	$5{,}92 \cdot 10^{-2}$	$1{,}01 \cdot 10^{-2}$
12.	Timur	186	3	1	$5{,}59 \cdot 10^{-2}$	$1{,}01 \cdot 10^{-2}$
13.	wiswayahe	23	1	1	$7{,}18 \cdot 10^{-3}$	$5{,}05 \cdot 10^{-3}$

Langkah keempat yaitu mengalikan peluang kemunculan masing-masing kategori data dengan peluang kemunculan setiap *term* pada data berkategori yang sama. Perhitungan pertama yaitu mengalikan peluang kemunculan data berkategori positif dengan peluang kemunculan setiap *term* pada dokumen positif dan perhitungan yang kedua mengalikan kemunculan data berkategori negatif dengan peluang kemunculan setiap *term* pada dokumen negatif. Perhitungan dari keduanya yaitu seperti berikut ini:

$$P(positif|contoh)$$
$$= p(positif) \cdot p("pkh"|positif) \cdot p("plus"|positif)$$
$$\cdot \ p("lansia"|positif) \cdot \ldots \cdot p("wiswayahe"|positif)$$
$$= 1{,}73 \cdot 10^{-27}$$

$$P(negatif|contoh)$$
$$= p(negatif) \cdot p("pkh"|negatif) \cdot p("plus"|negatif)$$
$$\cdot p("lansia"|negatif) \cdot \ldots \cdot p("wiswayahe"|negatif)$$
$$= 3{,}73 \cdot 10^{-30}$$

Dari dua perhitungan peluang di atas, terlihat bahwa peluang kemunculan kategori positif untuk contoh pertama mempunyai nilai yang lebih besar daripada negatif, oleh karena itu contoh pertama tersebut mempunyai kategori positif.

Contoh kedua yaitu *tweet* dengan isi "Pendukung Fanatik Khofifah Pindah Dukung Gus Ipul via @bhirawaonline" yang dalam proses *testing* masuk dalam kategori negatif.

Untuk membuat perhitungan *naive bayes* secara manual, tahap pertama yang harus dilakukan yaitu menghitung frekuensi kemunculan term pada contoh kedua pada data *training*. Tabel 4.9 menunjukkan frekuensi kemunculan *term* yang ada pada contoh kedua pada kumpulan data *training*.

Tabel 4.9 Tabel Frekuensi Kemunculan *Term* untuk Contoh Kedua

No	*Term*	Frekuensi Kategori *Training*	
		Positif	Negatif
1.	pendukung	13	6
2.	fanatik	0	2
3.	khofifah	1429	75
4.	pindah	0	2
5.	dukung	116	22
6.	Gus	140	51
7.	Ipul	47	43
8.	Bhirawaonline	0	1

Selain perhitungan jumlah *term* yang ada pada pada Tabel 4.9, harus juga diketahui jumlah semua term yang ada pada dokumen positif dan dokumen negatif. Pada dokumen positif terdapat 3343 *term* dan pada dokumen negatif terdapat 396 *term*.

Langkah kedua yaitu menghitung peluang kemunculan masing-masing kelas. Peluang kemunculan *tweet* berkategori positif adalah seperti perhitungan berikut ini:

$$P(positif) = \frac{Jumlah\ data\ positif}{Jumlah\ semua\ data}$$

$$P(positif) = \frac{1906}{2000}$$

$$P(positif) = 9{,}53 \cdot 10^{-1}$$

Sedangkan untuk kategori kelas negatif, peluang kemunculannya adalah sebagai berikut:

$$P(negatif) = \frac{Jumlah\ data\ negatif}{Jumlah\ semua\ data}$$

$$P(negatif) = \frac{94}{2000}$$

$$P(negatif) = 4{,}70 \cdot 10^{-2}$$

Langkah ketiga yaitu menghitung peluang kemunculan *term* pada masing-masing kategori dokumen. Berikut ini adalah contoh perhitungan peluang setiap *term* pada setiap kategori dokumen:

1. *term* "pendukung"

$$P(\text{“}pendukung\text{"}|positif) = \frac{n_k + 1}{n_{positif}} = \frac{13 + 1}{3343} = \frac{14}{3343} = 4{,}19 \cdot 10^{-3}$$

$$P(\text{“}pendukung\text{"}|negatif) = \frac{n_k + 1}{n_{negatif}} = \frac{6 + 1}{396} = \frac{7}{396} = 1{,}77 \cdot 10^{-2}$$

2. *term* "fanatik"

$$P(\text{“}fanatik\text{"}|positif) = \frac{n_k + 1}{n_{positif}} = \frac{0 + 1}{3343} = \frac{1}{3343} = 2{,}99 \cdot 10^{-4}$$

$$P(\text{“}fanatik\text{"}|negatif) = \frac{n_k + 1}{n_{negatif}} = \frac{2 + 1}{396} = \frac{3}{396} = 7{,}58 \cdot 10^{-3}$$

3. *term* "khofifah"

$$P(\text{“}khofifah\text{"}|positif) = \frac{n_k + 1}{n_{positif}} = \frac{1429 + 1}{3343} = \frac{1430}{3343} = 4{,}28 \cdot 10^{-1}$$

$$P(\text{“}khofifah\text{"}|negatif) = \frac{n_k + 1}{n_{negatif}} = \frac{75 + 1}{396} = \frac{76}{396} = 1{,}92 \cdot 10^{-1}$$

Dan seterusnya.

Tabel 4.10 menunjukkan hasil lengkap perhitungan manual langkah ketiga.

Tabel 4.10 Hasil Perhitungan Manual untuk Contoh Kedua

No	*Term*	Frekuensi Kategori *Training*		Frekuensi Dokumen Uji	Peluang Kemunculan *Term*	
		Positif	Negatif		Positif	Negatif
1.	Pendukung	13	6	1	$4{,}19 \cdot 10^{-3}$	$1{,}77 \cdot 10^{-2}$
2.	Fanatik	0	2	1	$2{,}99 \cdot 10^{-4}$	$7{,}58 \cdot 10^{-3}$
3.	Khofifah	1429	75	1	$4{,}28 \cdot 10^{-1}$	$1{,}92 \cdot 10^{-1}$
4.	Pindah	0	2	1	$2{,}99 \cdot 10^{-4}$	$7{,}58 \cdot 10^{-3}$
5.	Dukung	116	22	1	$3{,}50 \cdot 10^{-2}$	$5{,}81 \cdot 10^{-2}$
6.	Gus	140	51	1	$4{,}22 \cdot 10^{-2}$	$1{,}31 \cdot 10^{-1}$
7.	Ipul	47	43	1	$1{,}44 \cdot 10^{-2}$	$1{,}11 \cdot 10^{-1}$
8.	Bhirawaonline	0	1	1	$2{,}99 \cdot 10^{-4}$	$5{,}05 \cdot 10^{-3}$

Langkah keempat yaitu mengalikan peluang kemunculan masing-masing kategori data dengan peluang kemunculan setiap *term* pada data berkategori yang sama. Yang pertama yaitu mengalikan peluang kemunculan data berkategori positif dengan peluang kemunculan setiap *term* pada dokumen positif seperti berikut ini:

$$P(positif|contoh)$$
$$= p(positif) \cdot p("pendukung"|positif) \cdot p("fanatik"|positif)$$
$$\cdot p("khofifah"|positif) \cdot \ldots \cdot p("wiswayahe"|positif)$$
$$= 9{,}69 \cdot 10^{-19}$$

$$P(negatif|contoh)$$
$$= p(negatif) \cdot p("pendukung"|negatif)$$
$$\cdot p("fanatik"|negatif) \cdot p("khofifah"|negatif) \cdot \ldots$$
$$\cdot p("wiswayahe"|negatif)$$
$$= 3{,}92 \cdot 10^{-14}$$

Dari dua perhitungan peluang di atas, terlihat bahwa peluang kemunculan kategori negatif untuk contoh kedua mempunyai nilai yang lebih besar daripada positif, oleh karena itu contoh kedua tersebut mempunyai kategori negatif.

G. Evaluasi Sistem

Setelah sistem klasifikasi sudah selesai dibuat dari data *training* yang ada. Langkah selanjutnya yaitu dilakukan pengujian terhadap sistem klasifikasi yang dibuat. Evaluasi terhadap sistem ini merupakan tahap yang menentukan apakah sistem klasifikasi yang sudah dibuat mempunyai performa yang bagus atau tidak. Untuk melakukan pengujian terhadap sistem klasifikasi, biasanya digunakan *confusion matrix*. *Confusion matrix* ini bisa digunakan untuk menghitung akurasi, presisi, *recall*, dan *f-measure*. Untuk kasus klasifikasi yang mempunyai proporsi jumlah data yang tidak seimbang, akurasi tidak begitu mencerminkan performa dari sistem yang dibuat. Perhitungan presisi, *recall*, dan *f-measure* lebih penting untuk mengetahui performa dari sistem yang dibuat. Perhitungan evaluasi sistem klasifikasi dalam penelitian ini dibagi menjadi dua. Hal ini dikarenakan sistem yang dibuat ada dua yang masing-masing untuk calon gubernur yang ada.

Tabel 4.11 merupakan tabel *confusion matrix* untuk sistem pertama yang digunakan untuk menilai performa sistem.

Tabel 4.11 *Confusion Matrix* untuk Sistem Pertama

Prediksi / Aktual	Positif	Negatif	Total Baris
Positif	465	4	469
Negatif	1	27	28
Total Kolom	466	31	497

Berikut ini adalah hasil perhitungan akurasi, presisi, *recall*, dan *f-measure* dari sistem pertama berdasarkan Tabel 4.11:

$$Akurasi = \frac{465 + 27}{465 + 27 + 1 + 4} \cdot 100\%$$

$$= 98{,}99\%$$

$$Presisi\ Positif = \frac{465}{465 + 1} \cdot 100\%$$

$$= 99{,}78\%$$

$$Presisi\ Negatif = \frac{27}{27 + 4} \cdot 100\%$$

$$= 87{,}09\%$$

$$Rata - Rata\ Presisi = 93{,}44\%$$

$$Recall\ Positif = \frac{465}{465 + 4} \cdot 100\%$$

$$= 99{,}14\%$$

$$Recall\ Negatif = \frac{27}{27 + 1} \cdot 100\%$$

$$= 96{,}42\%$$

$$Rata - rata\ Recall = 97{,}78\%$$

$$F - Measure = 2 \cdot (\frac{93{,}44\% \cdot 97{,}78\%}{93{,}44\% + 97{,}78\%})$$

$$= 95{,}56\%$$

Tabel 4.12 menunjukkan *confusion matrix* untuk sistem kedua yang digunakan untuk menilai performa sistem.

Tabel 4.12 *Confusion Matrix* untuk Sistem Kedua

Prediksi / Aktual	Positif	Negatif	Total Baris
Positif	236	2	238
Negatif	1	48	49
Total Kolom	237	50	287

Berdasarkan Tabel 4.12, performa sistem yang kedua bisa dinilai melalui perhitungan akurasi, presisi, *recall*, dan *f-measure* berikut ini:

$$Akurasi = \frac{236 + 48}{236 + 48 + 1 + 2} \cdot 100\%$$

$$= 98{,}95\%$$

$$Presisi\ Positif = \frac{236}{236 + 1} \cdot 100\%$$

$$= 99{,}57\%$$

$$Presisi\ Negatif = \frac{48}{48 + 2} \cdot 100\%$$

$$= 96\%$$

$$Rata - rata\ Presisi = 97{,}78\%$$

$$Recall\ Positif = \frac{236}{236 + 2} \cdot 100\%$$

$$= 99{,}15\%$$

$$Recall\ Negatif = \frac{48}{48+1} \cdot 100\%$$

$$= 97{,}95\%$$

$$Rata - rata\ Recall = 98{,}55\%$$

$$F - Measure = 2 \cdot (\frac{97{,}78\% \cdot 98{,}55\%}{97{,}78\% + 98{,}55\%})$$

$$= 98{,}17\%$$

H. Uji Coba Sistem

Setelah kedua sistem yang dibuat mempunyai nilai presisi, *recall*, dan *f-measure* yang bagus (lebih dari 85%). Maka langkah selanjutnya yaitu melakukan uji coba terhadap data yang lebih banyak, sehingga diketahui kecenderungan sentimen masyarakat terhadap masing-masing calon gubernur. Berikut ini adalah hasil uji coba sistem pertama dan kedua terhadap masing-masing data:

Tabel 4.13 Tabel Hasil Uji Coba Sistem

No	Data	Positif	Persentase Data Positif	Negatif	Persentase Data Negatif	Total
1	Data calon gubernur pertama	18521	92,86%	1423	7,13%	19944
2	Data calon gubernur kedua	13714	85,98%	2235	14,01%	15949
	Total	32235		3658		35893

Pada Tabel 4.13, terlihat bahwa data yang digunakan dalam penelitian ini berjumlah total 35893 data yang terbagi ke dalam dua data yaitu data pertama untuk

calon gubernur nomor urut satu dan data kedua untuk calon nomor urut dua. Data ini didapatkan dari hasil *crawling* data dengan rstudio pada tanggal 22 April 2018 sampai tanggal 7 Juni 2018. Data untuk calon gubernur nomor urut satu berjumlah 19944 data dan data untuk calon gubernur nomor urut dua berjumlah 15949 data.

Untuk data calon gubernur nomor urut satu, sistem yang telah dibuat mengklasifikasikan semua data yang ada menjadi 18521 data bersentimen positif dan 1423 data bersentimen negatif. Hal ini berarti bahwa dari data yang ada, calon gubernur nomor urut satu mendapatkan sentimen positif sebanyak

$$\frac{18521}{19944} \cdot 100\% = 92{,}86\%$$

dan mendapatkan sentimen negatif sebanyak

$$\frac{1423}{19944} \cdot 100\% = 7{,}13\%$$

Untuk data calon gubernur nomor urut dua, sistem yang telah dibuat mengklasifikasikan semua data yang ada menjadi 13714 data bersentimen positif dan 2235 data bersentimen negatif. Hal ini berarti bahwa dari data yang ada, calon gubernur nomor urut satu mendapatkan sentimen positif sebanyak

$$\frac{13714}{15949} \cdot 100\% = 85{,}98\%$$

Dan mendapatkan sentimen negatif sebanyak

$$\frac{2235}{15949} \cdot 100\% = 14{,}01\%$$

Dari gabungan data pertama dan data kedua yang didapatkan dari proses *crawling* twitter, dapat disimpulkan bahwa calon gubernur nomor urut satu lebih unggul daripada calon gubernur nomor urut dua. Kesimpulan ini dapat diambil berdasarkan fakta bahwa:

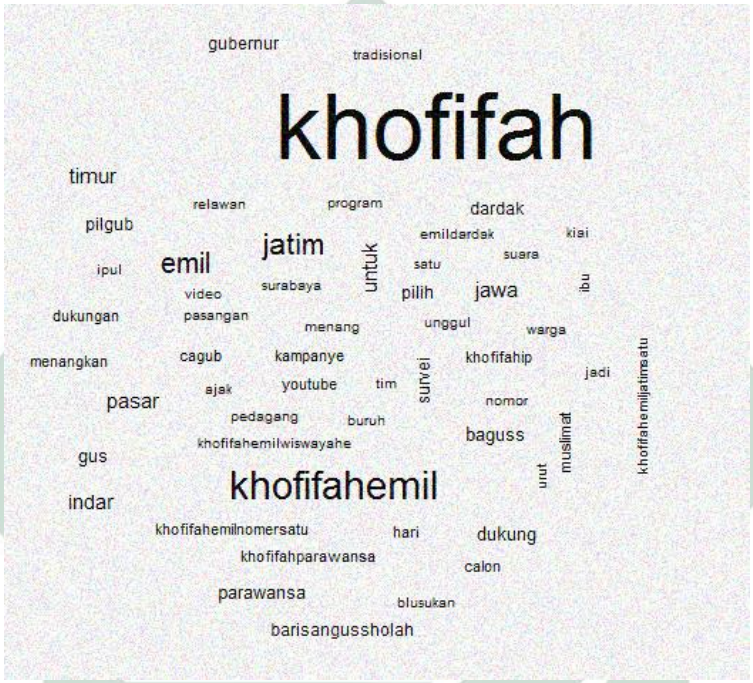
1. Calon gubernur nomor urut satu mendapatkan perhatian yang lebih banyak dari para pengguna twitter dengan bukti lebih banyaknya data yang menyebutkan calon gubernur nomor urut satu.
2. Persentase sentimen positif terhadap calon gubernur nomor urut satu lebih banyak daripada calon gubernur nomor urut dua dan persentase sentimen negatif terhadap calon gubernur nomor urut satu lebih sedikit daripada calon gubernur nomor urut dua.

I. Visualisasi dan Analisis Hasil

Visualisasi dan analisis hasil dalam penelitian ini dibagi menjadi tiga tahap. Tahap yang pertama yaitu menggunakan grafik *wordcloud* yang berasal dari software Rstudio. Grafik *wordcloud* ini merupakan grafik yang menunjukkan frekuensi kata dalam suatu dokumen. Tahap yang kedua yaitu menampilkan analisis hubungan antar kata yang dibuat dengan software VOSviewer. Dalam VOSviewer, visualisasi dari hasil ditampilkan dalam bentuk *network graph*. *Network graph* merupakan sebuah grafik yang menggambarkan hubungan antara kata-kata yang mempunyai frekuensi yang tinggi pada dokumen tersebut. Dari *network graph tersebut*, bisa dianalisis hubungan yang ada pada kata-kata yang ada pada dokumen tersebut. Sedangkan tahap yang ketiga menampilkan perbandingan jumlah *tweet*

masing-masing calon dan lima akun masing-masing calon gubernur yang jumlah *retweet* terbanyak.

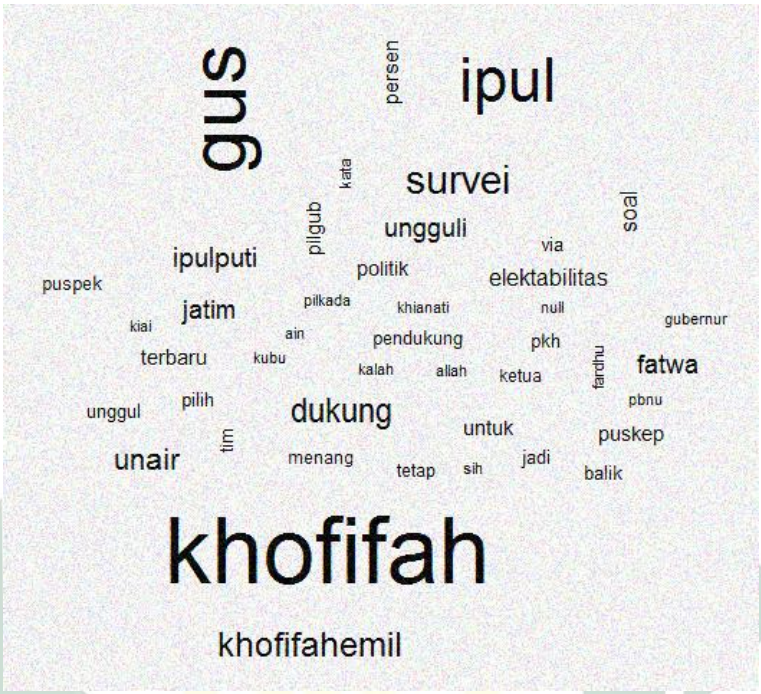
Gambar 4.3 adalah tampilan *wordcloud* untuk data sistem pertama yang berkategori positif dengan frekuensi minimal 50 kali:



Gambar 4.3 *Wordcloud* untuk Data Sistem Pertama yang Berkategori Positif

Dari Gambar 4.3, terlihat bahwa kata yang mempunyai frekuensi yang tinggi pada data sistem pertama adalah nama-nama yang sering digunakan untuk menyebutkan calon gubernur nomor urut satu dan pasangannya seperti khofifah, khofifahemil, khofifahparawansa, emildardak, dll. Setelah itu, terlihat juga slogan-slogan yang sering digunakan para pendukung calon gubernur nomor urut satu seperti khofifahemilwiswayahe, khofifahemilnomorsatu, khofifahemiljatimsatu, dll. Hal ini menunjukkan bahwa kata-kata yang tersebut dalam penelitian ini mempunyai sentimen yang cenderung positif ke calon gubernur nomor urut satu.

Gambar 4.4 adalah tampilan *wordcloud* untuk data sistem pertama yang berkategori negatif dengan frekuensi minimal 5 kali:

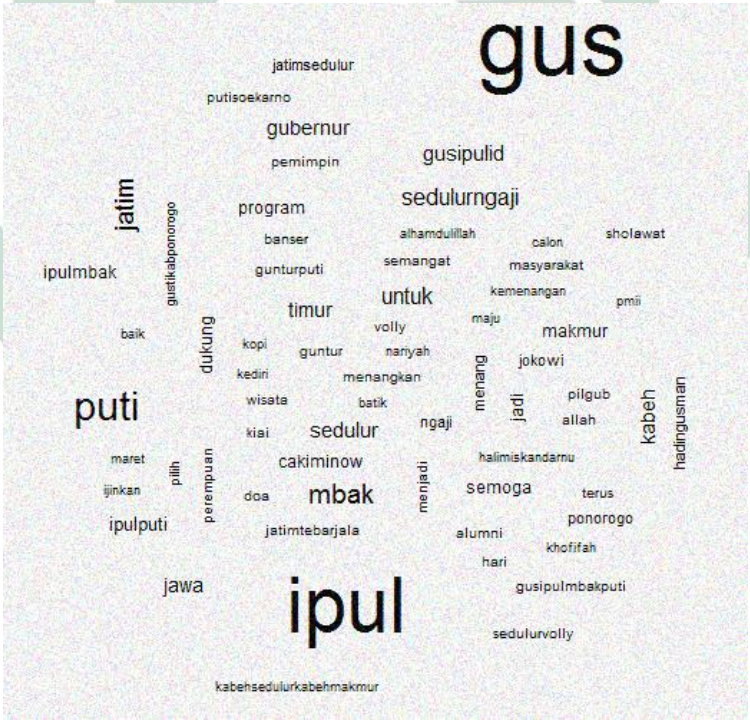


Gambar 4.4 *Wordcloud* untuk Data Sistem Pertama yang Berkategori Negatif

Gambar 4.4 menunjukkan bahwa kata-kata yang mempunyai frekuensi tinggi adalah khofifah, gus ipul, puskep, pkh, fatwa, dll. Hal ini mengindikasikan bahwa kombinasi dari kata-kata tersebut cenderung menghasilkan data *tweet* yang mempunyai kategori negatif. *Tweet* bersentimen negatif untuk calon gubernur nomor urut satu sebagian besar fokus menyinggung tentang penyalahgunaan program PKH dan fatwa fardhu ain untuk memilih khofifah. PKH adalah singkatan dari Program Keluarga Harapan. Khofifah dituding melakukan penyalahgunaan PKH di Lamongan sebagai alat kampanye. Namun, hal ini tidak terbukti dan Panitia Pengawas Pemilu Lamongan sudah melakukan klarifikasi bahwa tidak ada pelanggaran seperti yang dituduhkan (Hardiyanto, 2018). Pada saat masa kampanye

pemilihan Kepala Daerah Jawa Timur 2018, tim pemenangan pasangan Khofifah-Emil mengeluarkan fatwa fardhu ain untuk memilih Khofifah-Emil (Ardlyanto, 2018). Hal ini digunakan pendukung pihak lawan untuk melancarkan sentimen negatif yang menyebutkan bahwa tim pemenangan khofifah gunakan segala cara untuk menang, fatwa fardhu ain tak berdasar, dan lain-lain (Ardlyanto, 2018).

Gambar 4.5 adalah tampilan *wordcloud* untuk data sistem kedua yang berkategori positif dengan frekuensi minimal 30 kali:

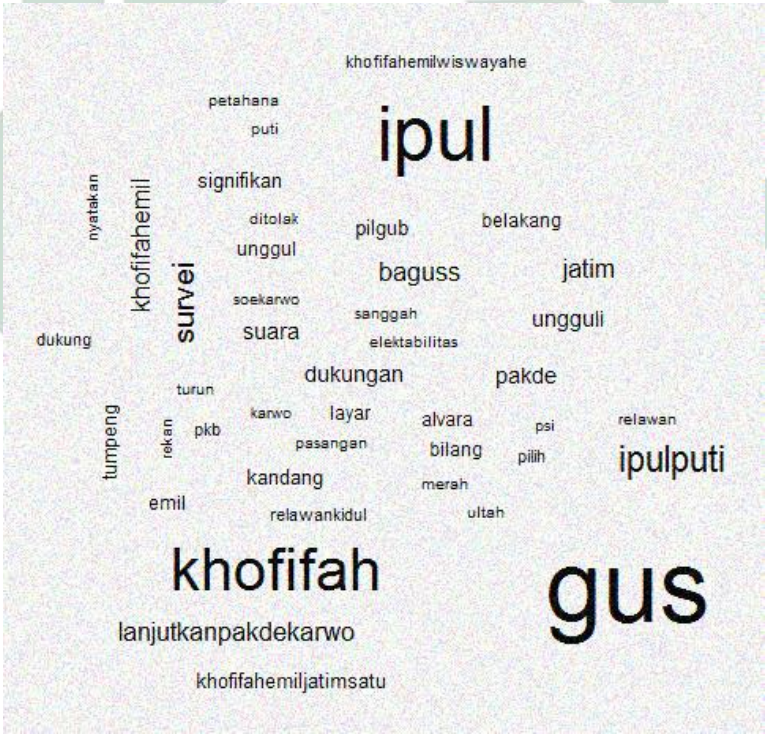


Gambar 4.5 *Wordcloud* untuk Data Sistem Kedua yang Berkategori Positif

Gambar 4.5 menunjukkan kata-kata yang mempunyai frekuensi kemunculan yang tinggi pada data sistem kedua yag berkategori positif. Kata-kata yang sering muncul adalah gus ipul, mbak puti, dan slogan-slogan pendukung pasangan calon gubernur nomor urut dua seperti kabehsedulurkabehmakmur, sedulurjatim,

jatimtebarjala, dll. Jadi, kemunculan slogan-slogan di atas merupakan indikasi bahwa *tweet* yang bersangkutan mempunyai kategori positif. Jatim tebar jala (Pusat ekonomi baru jalur selatan) merupakan program memakmurkan wilayah selatan Jawa Timur yang selama ini kalah dari wilayah utara Jawa Timur untuk mengurangi ketimpangan diantara keduanya..

Gambar 4.6 adalah tampilan *wordcloud* untuk data sistem kedua yang berkategori negatif dengan frekuensi minimal 5 kali:

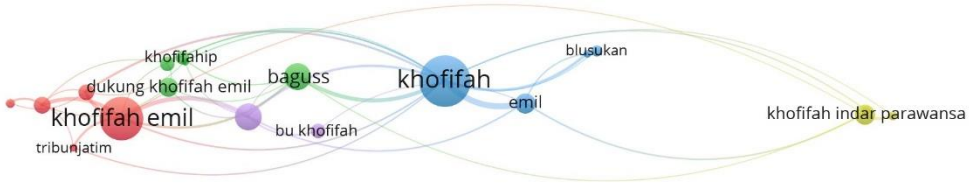


Gambar 4.6 *Wordcloud* untuk Data Sistem Kedua yang Berkategori Negatif

Pada Gambar 4.6, terlihat bahwa kata-kata seperti lanjutkanpakdekarwo, khofifahemiljatimsatu, khofifahemilwiswayahe, tumpeng ketika disandingkan dengan kata gus ipul atau mbak puti, maka kemungkinan besar *tweet* tersebut mempunyai kategori sentimen negatif terhadap calon gubernur nomor urut dua.

Lanjutkanpakdekarwo adalah *hashtag* yang digunakan pendukung setia Soekarwo, petahana gubernur Jawa Timur yang juga merupakan Ketua DPP Partai Demokrat Jawa Timur. Partai Demokrat juga merupakan partai pengusung pasangan Khofifah-Emil dalam pemilihan Kepala Daerah Jawa Timur 2018 sehingga *hashtag* lanjutkanpakdekarwo sebagian besar mempunyai sentimen positif kepada pasangan nomor satu atau sentimen negatif kepada pasangan nomor urut dua. Kata tumpeng ketika muncul bersama dengan gus ipul kemungkinan besar mempunyai sentimen negatif. Hal ini terkait dengan kabar tim pemenangan gus Ipul menolak tumpeng persaudaraan dari Khofifah. Menurut ketua tim pemenangan gus Ipul-mbak Puti, Hikmah Bafaqih, tumpeng ulang tahun dari Khofifah sudah diterima dengan baik di posko pemenangan jalan Gayungsari (Witanto, 2018).

Gambar 4.7 menampilkan *network graph* dari data untuk calon gubernur dengan nomor urut satu yaitu Khofifah Indar Parawansa:



Gambar 4.7 Hasil Visualisasi Data Khofifah Indar Parawansa

Gambar 4.7 memvisualisasikan hubungan yang ada pada data sistem pertama. Berikut ini adalah hubungan yang divisualisasikan oleh Gambar 4.7 tersebut:

Pada *cluster* pertama yang berwarna merah, ada lima frasa yang berhubungan satu sama lain yaitu khofifah emil unggul, gus ipul puti, bagussholah, khofifah emil, dan tribun jatim. Hubungan yang ada dari frasa-frasa tersebut adalah khofifah emil unggul atas gus ipul puti yang diberitakan oleh tribun jatim. Selain itu, adanya frasa bagussholah sendiri maksudnya adalah kebanyakan *tweet* yang memberitakan berita keunggulan khofifah emil atas gus ipul puti ditulis oleh barisan gus sholah yang merupakan pendukung dari pasangan khofifah dan emil.

Pada *cluster* kedua yang berwarna hijau, terdapat empat frasa yang berhubungan satu sama lain yaitu khofifahip, emildardak, wiswayahe, dukung khofifah emil, dan baguss. Hubungan yang ada dari frasa-frasa tersebut adalah baguss atau singkatan dari barisan gus sholah mendukung pasangan khofifah dan emil dengan semboyan khasnya yaitu wiswayahe. Adapun khofifahip dan emildardak merupakan akun twitter resmi dari Khofifah Indar Parawansa dan Emil Elestianto Dardak.

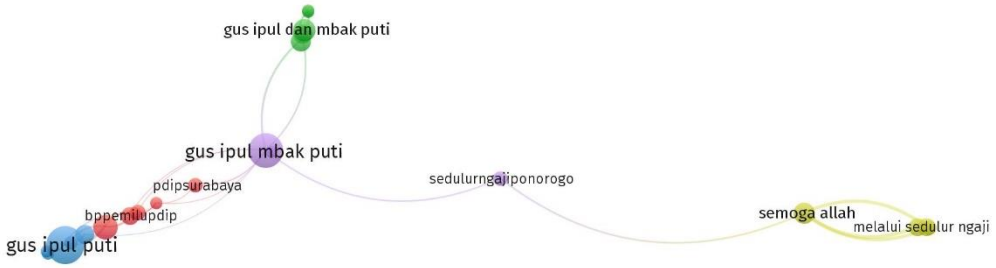
Pada *cluster* ketiga yang berwarna ungu, terdapat dua frasa yaitu bu khofifah dan barisangussholah. Hubungan yang ada dari frasa-frasa tersebut yaitu barisan gus sholah merupakan pendukung dari bu khofifah.

Pada *cluster* keempat yang berwarna biru, terdapat empat frasa yaitu khofifah, emil, blusukan, dan aliansi santri pemuda ekonom kiai. Hubungan yang ada dari frasa-frasa tersebut adalah aliansi santri pemuda ekonom kiai merupakan salah satu pendukung Khofifah Indar Parawansa untuk mencalonkan diri sebagai gubernur

Jawa Timur. Selain itu, blusukan merupakan rutinitas pasangan khofifah dan emil jelang pelaksanaan pemilihan gubernur.

Pada *cluster* kelima yang berwarna kuning, terdapat tiga frasa yaitu khofifah indar parawansa, calon gubernur jawa timur nomor urut, dan cagub jawa timur nomor urut. Hal ini mengindikasikan bahwa ketiga frasa tersebut merupakan frasa yang sering disebut beriringan karena ketiganya merupakan hal yang sama.

Dari kelima *cluster* terlihat bahwa kata khofifah merupakan kata yang mempunyai frekuensi paling tinggi dan merupakan pusat dari semua *cluster* karena dari semua *cluster* berhubungan dengan khofifah.



Gambar 4.8 Hasil Visualisasi Data Saifullah Yusuf

Gambar 4.8 memvisualisasikan hubungan antar kata atau frasa yang ada pada data sistem kedua. Data sistem kedua terbagi ke dalam lima *cluster*. Berikut ini adalah hubungan yang divisualisasikan oleh Gambar 4.8 tersebut:

Pada *cluster* pertama yang berwarna biru, terdapat lima frasa yang saling berhubungan yaitu detikcom, survei, gus ipul puti, gus ipul menang, seribu dewi gus ipul puti. Hal ini mengindikasikan bahwa menurut survei yang diberitakan leh detikcom, gus ipul unggul dibandingkan khofifah. Selain itu, hal ini juga berarti

detikcom banyak memberitakan tentang gus ipul dengan program andalan yaitu seribu dewi atau seribu desa wisata.

Pada *cluster* kedua yang berwarna merah, terdapat enam frasa yang saling berhubungan yaitu gusipul mbak puti, bp pemilu pdip, puti guntur fighter, kabeh sedulur kabeh makmur, gus ipul mbak puti menang, dan pdipsurabaya. Hal ini mengindikasikan bahwa pdip melalui bp pemilu mendukung penuh pasangan gus ipul puti. Selain itu, menurut pendukung pasangan gus ipul mbak puti yang mempunyai semboyan kabeh sedulur kabeh makmur, puti merupakan seorang fighter atau pejuang yang membuat para pendukung optimis akan kemenangan pada pemilihan gubernur.

Pada *cluster* ketiga yang berwarna ungu, terdapat tiga frasa yang saling berhubungan yaitu gus ipul puti, tidak punya rasa ego, dan sedulur ngaji ponorogo. Hal ini mengindikasikan bahwa menurut pendukung gus ipul dan mbak puti yang salah satunya adalah sedulur ngaji ponorogo, pasangan ini tidak memiliki ego sehingga merupakan calon pemimpin yang baik.

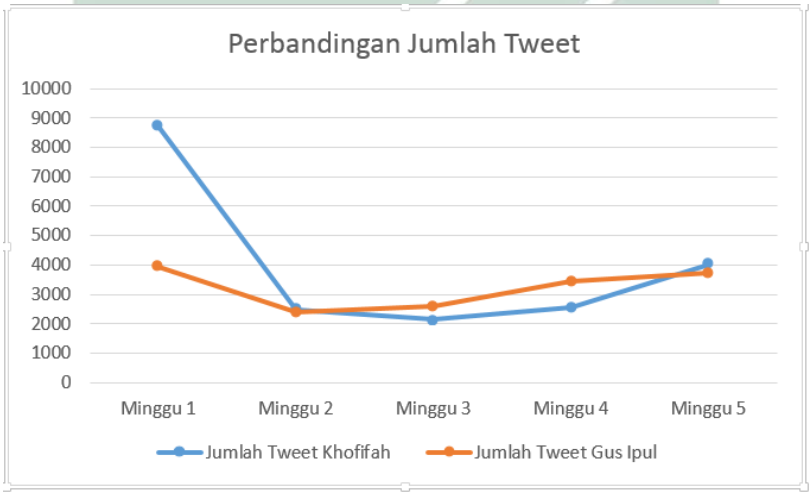
Pada *cluster* keempat yang berwarna hijau, terdapat enam frasa yang saling berhubungan yaitu menggalang semangat untuk memenangkan gus ipul, final sedulur volly hari minggu, gus ipul dan mbak puti, sedulur volly jokowi, dukung gus ipul, dan cakiminow. Hal ini menunjukkan bahwa final sedulur volly diadakan untuk menggalang semangat untuk memenangkan gus ipul. Selain itu, para pendukung setia gus ipul melakukan klaim bahwa jokowi merupakan salah satu tokoh nasional yang mendukung gus ipul untuk menjadi gubernur jawa timur. Di

lain sisi, pada *tweet* tentang gus ipul, terdapat banyak pendukung gus ipul yang melakukan *mention* terhadap akun twitter Muhaimin Iskandar yang merupakan ketua umum partai kebangkitan bangsa yang merupakan partai pengusung gus ipul.

Pada *cluster* kelima yang berwarna kuning, terdapat empat frasa yang saling berhubungan yaitu semoga Allah, melalui sedulur ngaji, sedulur ngaji kota batu, gus ipul mbak puti jadi gubernur dan wagub jatim. Hal ini mengindikasikan bahwa para pendukung gus ipul melalui sedulur ngaji berharap semoga Allah memberikan kemenangan kepada gus ipul dan mbak puti.

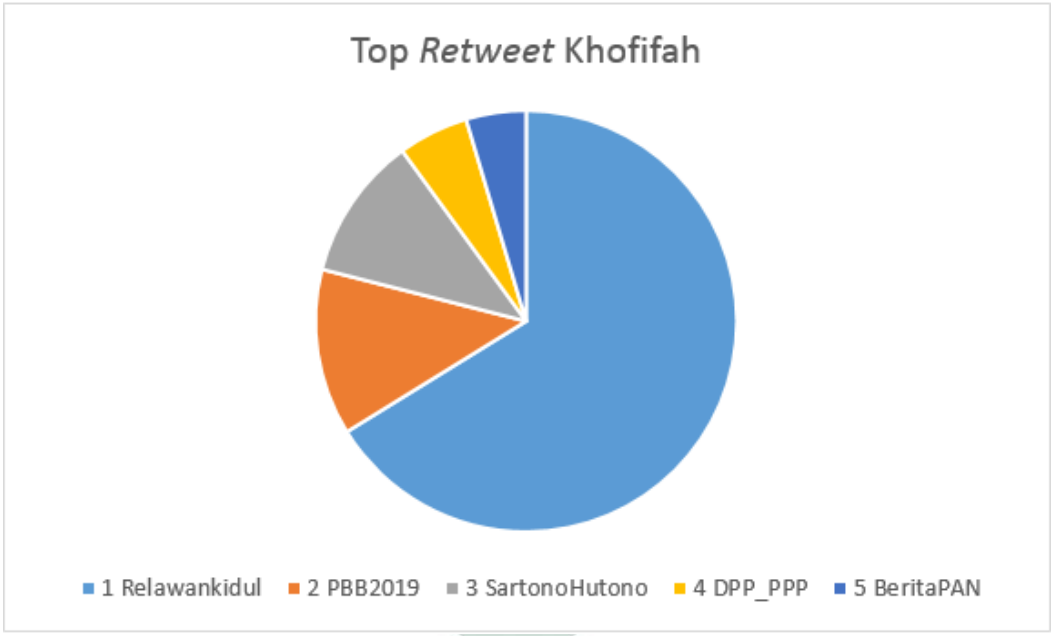
Dari lima *cluster* pada Gambar 4.4, terlihat bahwa pusat semua *cluster* adalah frasa gus ipul mbak puti yang mempunyai hubungan dengan empat *cluster* yang lain dan juga frasa tersebut mempunyai frekuensi kemunculan paling tinggi.

Berikut ini adalah perbandingan jumlah *tweet* yang menyebutkan Khofifah dengan *tweet* yang menyebutkan Gus Ipul:



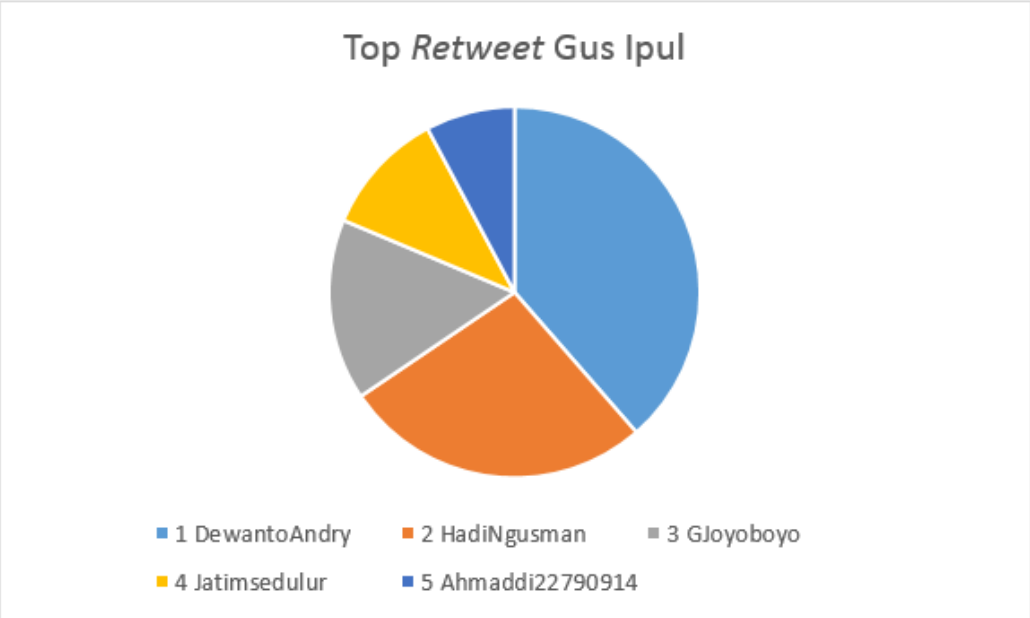
Gambar 4.9 Perbandingan Jumlah *Tweet* tentang Khofifah dengan Gus Ipul

Gambar 4.9 menunjukkan bahwa selama pengambilan data twitter, banyak data yang didapat hampir sama kecuali data pada minggu pertama yang jumlah *tweet* tentang Khofifah berjumlah lebih dari dua kali jumlah *tweet* tentang Gus Ipul. Hal ini tentu membuat jumlah *tweet* tentang Khofifah mempunyai jumlah yang lebih banyak daripada *tweet* tentang Gus Ipul. Hal ini juga membuktikan bahwa masyarakat pada media sosial twitter lebih cenderung memilih Khofifah dibandingkan dengan Gus Ipul.



Gambar 4.10 Peringkat 5 Besar Akun Pendukung Khofifah

Dari Gambar 4.10, terlihat bahwa urutan akun pendukung Khofifah dengan jumlah *retweet* terbanyak yaitu pertama @Relawankidul dengan jumlah 576 *retweet*, kedua @PBB2019 dengan jumlah 111 *retweet*, ketiga @SartonoHutomo dengan jumlah 96 *retweet*, keempat @DPP_PPP dengan jumlah 47 *retweet*, dan terakhir @BeritaPAN dengan jumlah 40 *retweet*.



Gambar 4.11 Peringkat 5 Besar Akun Pendukung Gus Ipul

Gambar 4.11 menunjukkan urutan akun pendukung Gus Ipul dengan jumlah *retweet* terbanyak. Urutan pertama yaitu @DewantoAndry dengan jumlah 149 *retweet*, kedua yaitu @HadiNgusman dengan jumlah *retweet* sebanyak 104 retweet, ketiga yaitu @Gjoyoboyo dengan jumlah 61 *retweet*, keempat yaitu @Jatimsedulur dengan jumlah 42 *retweet*, dan terakhir yaitu @Ahmaddi22790914 dengan jumlah 30 *retweet*.

BAB V

PENUTUP

A. Simpulan

Berdasarkan hasil pembuatan sistem klasifikasi dan uji coba sistem dengan menggunakan *naive bayes classifier* terhadap semua data yang diperoleh dari twitter, dapat disimpulkan bahwa:

1. *Naive Bayes Classifier* merupakan metode yang dapat diterapkan untuk melakukan klasifikasi sentimen dokumen twitter. Dalam penelitian ini, kasus yang digunakan yaitu opini pemilihan Kepala Daerah Jawa Timur 2018. Hal ini ditunjukkan dengan cukup tingginya performa sistem yang dibuat. Pada sistem yang pertama, didapatkan performa sistem yaitu sebagai berikut: akurasi sebesar 98,99%, presisi sebesar 93,44%, *recall* sebesar 97,78%, dan *f-measure* sebesar 95,56%. Sedangkan pada sistem yang kedua, didapatkan performa sistem sebagai berikut: akurasi sebesar 98,95%, presisi sebesar 97,78%, *recall* sebesar 98,55%, dan *f-measure* sebesar 98,17%.
2. Berdasarkan data yang diperoleh dari twitter, masyarakat Jawa Timur cenderung lebih memilih calon gubernur nomor urut satu yaitu Khofifah Indar Parawansa. Hal ini dibuktikan dengan jumlah *tweet* bersentimen positif terhadap calon gubernur nomor urut satu lebih banyak daripada calon gubernur nomor urut dua dengan jumlah 18521 *tweet* berbanding 13714 *tweet* dan jumlah *tweet* bersentimen negatif terhadap calon gubernur

nomor urut satu lebih sedikit daripada calon gubernur nomor urut dua dengan jumlah 1423 *tweet* berbanding 2235 *tweet*.

B. Saran

Penelitian ini masih terdapat banyak kekurangan. Oleh karena itu, penulis menyarankan beberapa hal berikut sebagai bahan perbaikan untuk penelitian dan pengembangan di masa depan:

1. Pada proses *crawling* data twitter, data yang diperoleh dari kata kunci "khofifah" banyak terdapat data outlier. Hal ini dikarenakan hasil *crawling* dengan kata kunci "khofifah" juga mendapatkan semua *tweet* dengan *username* yang mengandung kata "khofifah" meskipun *tweet* tersebut tidak berhubungan dengan calon gubernur nomor satu, Khofifah Indar Parawansa. Oleh karena itu, *crawling* sebaiknya dilakukan dengan kata kunci yang lebih tepat sehingga tidak didapatkan banyak data *outlier*.
2. Pada proses *crawling* didapatkan beberapa *tweet* yang terkesan terpotong dan tidak lengkap. Oleh karena itu, proses *crawling* data sebaiknya dilakukan dengan mencoba cara yang lain. Cara yang bisa dicoba adalah *crawling* dengan menggunakan bahasa pemrograman python atau dengan cara meminta data secara langsung kepada twitter melalui *Historical Data Power Track*.

DAFTAR PUSTAKA

Abdillah, S. (2015). *Penerapan Algoritma Decision Tree C4.5 untuk Diagnosa Penyakit Stroke dengan Klasifikasi Data Mining pada Rumah Sakit Santa Maria Pemalang.* Semarang: Universitas Dian Nuswantoro.

Aggarwal, C., & Zhai, C. (2013). A Survey of Text Classification Algorithm. *Mining Text Data*, 169-170.

Ardlyanto, A. (2018). *Panik, Tim Khofifah-Emil Keluarkan Fatwa Fardhu 'Ain*. Diakses pada 20 Juni 2018, dari merdeka.com: https:/www.merdeka.com

Balagatabi, Z. N. (2012). Comparison of Decision Tree and Naive Bayes Methods in Classification of Researcher's Cognitive Style in Academic Environment. *Journal of Advance in Computer Research, 3*(2), 23-24.

Buntoro, G. A. (2016). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 di Twitter. *Integer Journal, 1*(1), 32-41.

Dewanti, J. A. (2016). *Analyzing Public Sentiment Towards Public Policy on Social Media Twitter (Case Study: Ujian Nasional 2015 Policy).* Surabaya: Institut Teknologi Sepuluh November.

Faradhillah, N. Y. (2016). *Analisis Sentimen Terhadap Kinerja Pelayanan Publik di Kota Surabaya Berdasarkan Klasifikasi Komentar di Media Sosial dengan Menggunakan Naive Bayes.* Surabaya: Institut Teknologi Sepuluh November.

Hardiyanto, S. (2018). *Panwaslu Lamongan Pastikan Khofifah-Emil Tidak Memanfaatkan PKH*. Diakses pada 20 Juni 2018, dari JawaPos.com: https://www.jawapos.com/

Hidayatullah, A. F., & SN, A. (2014). Analisis Sentimen dan Klasifikasi Kategori Publik terhadap Tokoh Publik pada Twitter. *Seminar Nasional Informatika* (hal. 115-122). Yogyakarta: UPN "Veteran".

Juniawan, I. (2009). *Klasifikasi Dokumen Teks Berbahasa Indonesia Menggunakan Minor Component Analysis.* Bogor: Institut Pertanian Bogor.

Lestari, A. R., Perdana, R. S., & Fauzi, M. A. (2017, Desember). Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes dan Pembobotan Emoji. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 1*(12), 1718-1724.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining* (1st ed.). Morgan & Claypool Publishers.

Manning, C. (2008). *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Maulana, I. (2012). *Istilah-Istilah Dasar Dalam Twitter (Pemula)*. Diakses pada 10 April 2018, dari Irfan Maulana's Journal: https://mazipanneh.wordpress.com

Perdana, R. S. (2017). *Pengukuran Akurasi Menggunakan Precision dan Recall*. Diakses pada 10 April 2018, dari Rizal Setya Perdana: http://www.rizalespe.com

Prasetyo, E. (2014). *DATA MINING Mengolah Data Menjadi Informasi Menggunakan Matlab.* Yogyakarta: ANDI.

Putri, R. K., & Mukhlash, I. (2013). Penerapan Algoritma Klasifikasi Berbasis Aturan Asosiasi untuk Data Meteorologi. *Jurnal Sains dan Seni POMITS, 1*(1), 1-6.

Rini, D. C., Farida, Y., & Puspitasari, D. (2016). Klasifikasi Menggunakan Metode Hybrid Bayessian-Neural Network (Studi Kasus: Identifikasi Virus Komputer). *JURNAL MATEMATIKA "MANTIK", 01*(02), 38-43.

Rodiyansyah, S. F., & Winarko, E. (2012). Klasifikasi Posting Twitter Kemacetan Kota Bandung Menggunakan Naive Bayes Classification. *IJCCS, 6*(1), 91-100.

Saputro, B. (2016). *Klasifikasi dan Pemetaan Posdata Tematik berbasis Masjid menggunakan Naive Bayes Classifier.* Malang: UIN Maulana Malik Ibrahim.

Witanto. (2018). *Posko Pemenangan Gus Ipul Tidak Pernah Tolak Tumpeng, Meski Tidak Tahu Maksudnya.* Diakses pada 20 Juni 2018, dari merdeka.com: https://www.merdeka.com

Wong, A. H., & Abednego, L. (2015). *Pengelompokan Dokumen Otomatis dengan menggunakan TFIDf Classifier, Naive Bayes Classifier, dan KNN*. Bandung: Program Studi Teknik Informatika - UNPAR.

Wongso, R., Luwinda, F. A., Trisnajaya, B. A., Rusli, O., & Rudi. (2017). News Article Text Classification in Indonesian Language. *2nd International Conference on Computer Science and Computational Intelligence* (hal. 137-143). Bali: Elsevier.

Yulian, E. (2018). Text Mining dengan K-Means Clustering pada Tema LGBT dalam Arsip Tweet Masyarakat Kota Bandung. *JURNAL MATEMATIKA "MANTIK", 04*(01), 53-58.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF-IDF, LSI, and Multi-words for text classification. *Expert System with Application, 38*(3), 2758-2765.