

**LAPORAN PENELITIAN
HIBAH INTERNAL**

**PEMBANGUNAN MODEL SISTEM ANALISIS SENTIMEN
UNTUK PENGUKURAN PENERIMAAN MASYARAKAT
TERHADAP KEBIJAKAN PUBLIK YANG DITERAPKAN
PEMERINTAH (STUDI KASUS PILKADA JAWA BARAT)**



PENELITI

Ir. Munawar, MMSI., M.Com, PhD

Ir. Nizirwan Anwar, MT

PROGRAM STUDI/JURUSAN TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS ESA UNGGUL

TAHUN 2018


HALAMAN PENGESAHAN

1. Judul Penelitian : Pembangunan Model Sistem Analisis Sentimen untuk Pengukuran Penerimaan Masyarakat terhadap Kebijakan Publik yang Diterapkan Pemerintah (Studi Kasus Pilkada Jawa Barat)
2. Ketua Peneliti
a. Nama lengkap dengan gelar : Ir. Munawar MMSI., M.Com., PhD
b. Pangkat/Gol/NIP :
c. Jabatan Fungsional/Struktural : Lektor Kepala
d. Program Studi/Jurusan : Teknik Informatika
e. Fakultas : Fasilkom
f. Alamat Rumah/HP : 08128100435
g. E-mail : an_moenawar@yahoo.com
- Anggota Peneliti
a. Nama lengkap dengan gelar : Ir. Nizirwan Anwar, MT
c. Jabatan Fungsional/Struktural : Lektor Kepala
d. Program Studi/Jurusan : Teknik Informatika
3. Jumlah Tim Peneliti : 2 orang
4. Lokasi Penelitian : Jakarta dan sekitarnya
5. Kerjasama (kalau ada)
a. Nama Instansi :-
b. Alamat :-
6. Jangka waktu penelitian : 12. bulan
7. Biaya Penelitian : Rp. 14.500.000,00 (Empat Belas Juta Lima Ratus Ribu Rupiah)

Jakarta, 19 Agustus 2018


Ketua Peneliti

Mengetahui
Dekan Fakultas Ilmu Komputer


Dr. Ir. Husni S Sastramihardja
NIK: 214030494


Ir. Munawar, MMSI., M.Com., PhD
NIK: 202080208

Menyetujui,
Ketua Lembaga Penelitian dan Pengabdian kepada Masyarakat
Universitas Esa Unggul


Dr. Hasyim, SE., MM., M.Ed
NIK. 201040164

DAFTAR ISI

HALAMAN SAMPUL	i
HALAMAN PENGESAHAN	ii
DAFTAR ISI	iii
RINGKASAN	iv
BAB 1. PENDAHULUAN.....	1
Bab 1.1. Latar Belakang	1
Bab 1.2. Tujuan Penelitian	2
Bab 1.3. Ruang Lingkup	2
Bab 1.4. Manfaat Penelitian	3
Bab 1.5. Metode Penelitian	3
Bab 1.5.1. Analisis Masalah	3
Bab 1.5.2. Pengumpulan Data	3
Bab 1.5.3. Proses Text Mining	4
Bab 1.5.4. Penerapan Metode Klasifikasi Naive Bayes Classifier	4
BAB 2. TINJAUAN PUSTAKA	5
Bab 2.1. Pengertian Analisis Sentimen	5
Bab 2.2. Tingkatan Analisis Sentimen	5
Bab 2.3. Teknik Klasifikasi	7
Bab 2.4. Proses Klasifikasi	8
Bab 2.5. Algoritma Klasifikasi	10
BAB 3. TAHAPAN PENELITIAN	12
BAB 4. HASIL DAN PEMBAHASAN	14
Bab 4.1. Pengumpulan Data Twitter	14
Bab 4.2. Pra Proses Data	14
Bab 4.3. Proses Klasifikasi	15
Bab 4.4. Pembahasan	16
BAB 5. KESIMPULAN DAN SARAN	18
Bab 5.1. Kesimpulan	18
Bab 5.2. Saran	18
DAFTAR PUSTAKA	19

RINGKASAN

Pengguna internet adalah salah satu konsumen terbesar dari suatu objek berita yang ditampilkan lewat media sosial maupun media online lainnya di internet. Ini menjadi potensi bagi sejumlah kalangan seperti lembaga survei dan penelitian hingga lembaga politik untuk mendapatkan data sentimen pengguna internet terhadap suatu objek masalah dalam hal ini adalah tokoh pilkada.

Teknik-teknik dalam laporan ini dikembangkan untuk memenuhi tujuan tersebut di atas dengan memanfaatkan beberapa algoritma *data mining* dengan teknik klasifikasi dari data media social yang diambil dengan mempergunakan layanan antarmuka pemrograman aplikasi yang telah disediakan media social. Data tersebut diproses dengan *text mining* untuk menghindari data yang kurang sempurna. Selanjutnya data tersebut diklasifikasi menjadi klasifikasi positif, negatif, dan netral. Dengan proses analisis sentimen, popularitas tokoh pilkada dapat diukur dan digambarkan secara visual.

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

1. PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi informasi yang pesat membuat pertukaran informasi dan komunikasi menjadi semakin mudah. Munculnya media sosial seperti *Twitter*, *Facebook*, *Yahoo*, *Google*, *Youtube*, *Instagram*, dan *Path* telah mengubah kampanye para tokoh yang akan berlaga di pilkada (pemilihan kepala daerah). Oleh karena itu perlu alat bantu yang tepat untuk menyajikan data yang sangat besar yang dipicu oleh adanya media sosial ini.

Media sosial merupakan salah satu media komunikasi populer saat ini. Hal ini terlihat dari peningkatan pengguna *twitter* yang tercatat di seluruh dunia. Berdasarkan data yang dirilis infografis (tahun 2014-2015) *Twitter* memiliki 302 juta pengguna aktif yang 80 persennya berasal dari perangkat mobile. Dari angka itu, 37 persen pengguna *Twitter* berusia 18-29 tahun, sedangkan 25 persen lainnya berada di rentan usia 30-49 tahun [2]. Dengan jumlah pengguna aktif sebanyak itu, *Twitter* menerima kicauan sebanyak 500 juta setiap harinya. Sebanyak 68 persen berupa kicauan balasan, 26 persen berupa kicauan, dan 6 persen adalah kicauan ulang [2].

Pengguna *twitter* yang semakin meningkat ini terlihat dari jutaan *tweets* yang di *posting* setiap harinya dengan berbagai topik yang berbeda. Data *tweets* ini dapat berupa persepsi publik baik ekonomi, perilaku sosial, fenomena alam, bahkan juga politik [1]. Pada Oktober 2013 saja, pengguna aktif *Twitter* di Indonesia mencapai 6,5% dan menempati urutan ketiga dari seluruh pengguna dunia setelah Amerika dan Jepang (<http://www.statista.com/topics/737/twitter/chart/1642/regional-breakdown-of-twitterusers/>, 30 Maret 2014). Dengan menggunakan asumsi prosentase di atas (6.5%) maka jumlah cuitan per harinya di Indonesia ada 32.500.000 cuitan. Data ini merupakan data yang sangat besar, yang jika bisa digunakan untuk memetakan sentimen seseorang atas suatu tokoh politik akan bisa menjadi masukan yang luar biasa bagi parpol pengusung tokoh politik yang akan berlaga di pilkada.

Analisis sentimen atau *opinion mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini [3]. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif.

Opinion mining bisa dianggap sebagai kombinasi antara *text mining* dan *natural language processing*. Salah satu metode dari *text mining* yang bisa digunakan untuk menyelesaikan masalah *opinion mining* adalah *Naïve Bayes Classifier (NBC)*. *NBC* bisa digunakan untuk mengklasifikasikan opini ke dalam opini positif dan negatif. *NBC* bisa berfungsi dengan baik sebagai metode pengklasifikasi teks.

Penelitian tentang penggunaan *NBC* sebagai metode pengklasifikasi teks telah dilakukan oleh SM Kamaruzzaman dan Chowdury Mofizur Rahman [4] serta Ashraf M Kibriya *et.al.* [5] pada tahun 2004. Dari proses pengujian secara kualitatif disebutkan bahwa teks bisa diklasifikasikan dengan akurasi yang tinggi.

Sedangkan dari *natural language processing*, salah satu metode yang bisa digunakan untuk menyelesaikan masalah opinion mining adalah *Part-of-Speech (POS) Tagging*. *POS Tagging* digunakan untuk memberikan kelas kata (*tag*) secara gramatikal ke setiap kata dalam suatu kalimat teks. Beberapa penelitian yang ditujukan untuk mengembangkan sistem *POS Tagging* dalam bahasa Indonesia, diantaranya dilakukan oleh Femphy Pisceldo *et.al.* pada tahun 2009 [6] menggunakan *Maximum Entropy* dan Alfan Farizki *et.al.* [7] pada tahun 2010 menggunakan *Hidden Markov Model*. Akurasi yang didapatkan berkisar antara 85% hingga 96%.

Berdasarkan penelitian yang telah ada sebelumnya, penelitian ini mencoba melakukan analisis sentimen data dengan mengklasifikasi data *twitter* berbahasa Indonesia pada tokoh politik yang sedang berlaga di pilkada. Data tersebut akan diproses dengan *text mining* untuk menghindari data yang kurang sempurna kemudian mengklasifikasi data *tweet* ke dalam tiga klasifikasi yaitu klasifikasi positif, negatif, netral. Klasifikasi ini menggunakan algoritma *Naïve Bayes Classifier*. Adapun rumusan permasalahan dalam penelitian ini “Bagaimana Menganalisis dan Mengklasifikasi Sentimen Pada Data Media Sosial Menggunakan *Text Mining* terhadap Tokoh Politik yang Sedang Berlaga di Pilkada?”

1.2. Tujuan Penelitian

Penelitian ini bertujuan untuk :

1. Membangun model analisis data dari media sosial twitter dengan algoritma *Naïve Bayes Classifier*.
2. Membuat rancang bangun sistem opinion mining dengan antar muka ke twitter

1.3. Ruang Lingkup

Agar model yang dikembangkan benar-benar mencerminkan kondisi riil, penelitian ini akan menggunakan data-data terkait dengan pilkada Jawa Barat dari media sosial twitter. Data-data tentang pilkada Jawa Barat dari semenjak kampanye akan dianalisis dengan menggunakan teknik ini. Selanjutnya hasil ini akan dibandingkan dengan hasil pilkada versi resmi KPU.

1.4. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat kepada pihak-pihak yang sedang berlaga di pilkada untuk menentukan bentuk kampanye yang lebih baik agar bisa menaikkan tingkat kepopuleran tokoh yang sedang berlaga di pilkada.

Hasil penelitian ini tidak hanya bisa digunakan untuk menilai kepopuleran seorang tokoh, namun bisa juga digunakan untuk menilai kepopuleran suatu produk atau jasa. Dengan demikian banyak manfaat yang bisa didapatkan dari hasil penelitian ini.

1.5. Metode Penelitian

Tahapan penelitian yang akan dilakukan mencakup hal-hal sebagai berikut:

- Analisis masalah
- Pengumpulan data
- Proses *Text Mining*
- Penerapan metode Klasifikasi *Naïve Bayes Classifier*

1.5.1. Analisis Masalah

Informasi yang terkandung di media sosial bisa digunakan untuk menilai popularitas seorang tokoh yang sedang berlaga di pilkada. Opini yang dikemukakan tentang seorang tokoh yang sedang berlaga di pilkada bisa positif atau negatif. Penilaian ini disebut dengan analisis sentimen.

Dengan membuat suatu alat bantu yang bisa menarik informasi yang terkandung di media sosial maupun media online lainnya tentang seorang tokoh yang sedang berlaga di pilkada secara otomatis, diharapkan akan bisa membuat proses penilaian popularitas seorang tokoh ini bisa dilakukan pula secara otomatis. Dengan demikian akan bisa membantu pihak-pihak yang berkepentingan untuk melakukan langkah-langkah yang dibutuhkan untuk menaikkan popularitas sang tokoh tersebut.

Hanya saja, untuk melakukan hal tersebut tidak bisa dilakukan dengan mudah. Opini yang dikeluarkan seseorang dalam bentuk teks lebih sering dalam bentuk tidak formal. Oleh karena itu dalam *text mining* ini akan melalui tahapan *preprocessing* dan Ekstraksi serta *naïve bayes classifier* sebagai metode klasifikasi nilai sentimen apakah opini bernilai positif atau negatif.

1.5.2. Pengumpulan Data

Pada penelitian ini pengumpulan data dimulai dengan penarikan data *tweet* dari *server twitter* dan kemudian disimpan kedalam database. Penarikan data *tweet* dilakukan dengan menggunakan fasilitas

Application Programming Interface (API) yang sudah disediakan mereka. *Application Programming Interface* (API) ini mengambil data kotor dari server *twitter*, yang selanjutnya akan difilter menjadi data bersih dan kemudian disimpan ke database.

1.5.3. Proses *Text Mining*

Pada proses *text mining* membutuhkan 2 (dua) tahapan yaitu :

1. *Preprocessing* ini dilakukan untuk menghindari data yang kurang sempurna, gangguan pada data, dan data-data yang tidak konsisten [11]
2. Ekstraksi Fitur

Tahapan pada *text preprocessing* yang dilakukan adalah:

- a. Melakukan *Cleaning* menghilangkan kata yang tidak diperlukan agar dokumen bersih.
- b. *Case folding* dengan mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf “a” sampai dengan “z” yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter[12].
- c. Tahap tokenizing / parsing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya.
- d. *Filtering* adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist/stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* adalah “yang”, “dan”, “di”, “dari”, dan seterusnya[4].

1.5.3. Penerapan metode klasifikasi *Naïve Bayes Classifier*

Klasifikasi pada penelitian ini menggunakan *Naïve Bayes Classifier*. Pada tahapan ini akan dilakukan klasifikasi dengan menggunakan algoritma *naïve bayes classiefier*. Algoritma ini akan menghitung nilai sentimen positif, negatif atau netral.

2. TINJAUAN PUSTAKA

2.1. Pengertian Analisis Sentimen

Menurut Liu (2008), *sentiment analysis* (analisis sentimen) atau sering disebut juga dengan *opinion mining* (penambangan opini) adalah studi komputasi untuk mengenali dan mengekspresikan opini, sentimen, evaluasi, sikap, emosi, subjektivitas, penilaian atau pandangan yang terdapat dalam suatu teks.

Dave et al (2003), menjelaskan bahwa sebuah alat bantu penambangan opini merupakan pemrosesan sekumpulan hasil pencarian dari suatu item yang diberikan, menghasilkan satu daftar atribut produk (misal kualitas, fitur, dan lain-lain) dan menghitung agregasi dari opini dari masing-masing atribut tersebut (rendah, sedang, tinggi).

Sentimen menurut Kamus Besar Bahasa Indonesia (KBBI) adalah:

1. pendapat atau pandangan yang didasarkan pada perasaan yang berlebih-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 hati; tidak senang; dendam.
4. reaksi yang tidak menguntungkan. Contoh: penurunan harga saham hanya disebabkan oleh *sentimen* pasar

Sedangkan opini menurut KBBI adalah pendapat atau pikiran atau pendirian.

2.2. Tingkatan Analisis Sentimen

Liu [8] membagi analisis dalam tiga tingkatan:

1. Tingkatan Dokumen

Pada tingkatan ini, analisis dilakukan menyeluruh terhadap satu dokumen untuk mengklasifikasikan apakah keseluruhan dokumen mengekspresikan sentimen positif atau negatif. Analisis hanya bisa dilakukan pada dokumen yang tidak membandingkan lebih dari satu entitas. Pada contoh tulisan di atas, ada lebih dari satu entitas yaitu *kinerja*, *sepak terjang*, dan *langkah*.

2. Tingkatan Kalimat

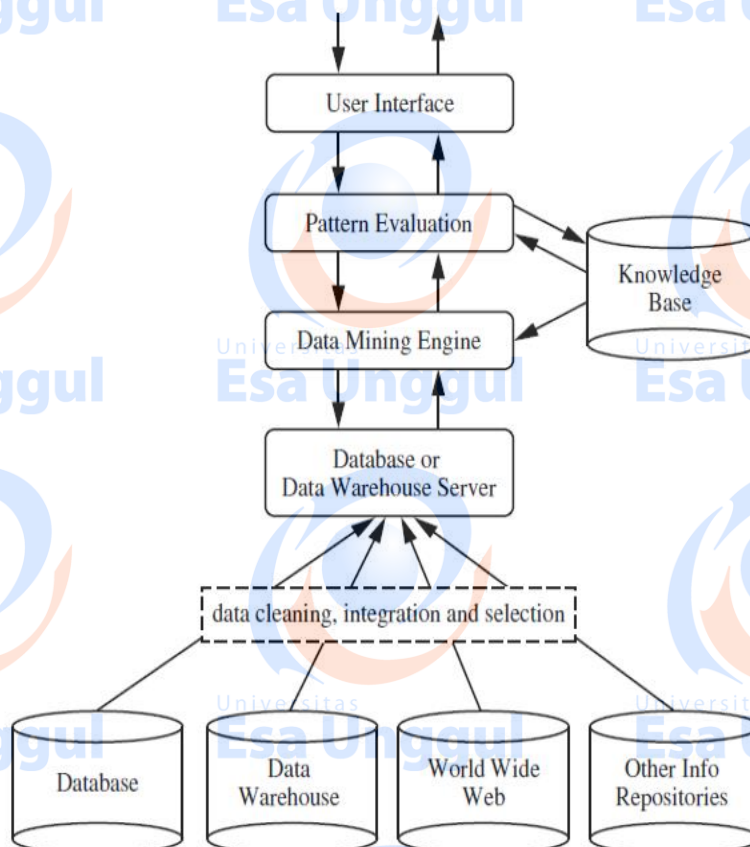
Pada tingkatan ini, analisis dilakukan pada kalimat untuk menentukan ekspresi tiap kalimat, apakah positif, negatif atau netral. Netral berarti tidak ada opini. Namun masih terdapat kendala untuk

membedakan mana fakta dan mana opini. Misal: “*Saya membeli iPhone bulan lalu, namun baterainya sudah rusak.*” Kalimat tersebut jelas fakta.

3. Tingkatan Entitas atau Aspek/Fitur

Kedua tingkatan sebelumnya ternyata sulit untuk menentukan apa yang sebenarnya orang suka dan tidak suka. Tingkatan aspek lebih bisa menentukan dengan pasti. Alih-alih melihat konstruksi bahasa (dokumen, paragraph, kalimat, klausa, frase), tingkatan aspek melihat langsung ke opini itu sendiri. Dengan ide dasar bahwa opini pasti punya sentimen dan punya target opini. Maka opini yang tidak terdapat target, tidak akan digunakan. Misal: “*Kinerja Jokowi memang bagus, namun janjinya diingkari.*” Ada dua aspek yaitu kinerja Jokowi dan janji Jokowi. Sentimen pada kinerja bernilai positif. Sentimen pada janji bernilai negatif. Kinerja Jokowi dan janji Jokowi adalah target opini. Pada tingkatan ini, ringkasan struktur dari opini tentang entitas atau aspek tertentu dapat dibuat.

Secara umum arsitektur data mining bisa digambarkan sebagai berikut:



Gambar 2.1. Arsitektur data mining secara umum [9]

Lapisan paling bawah merupakan satu atau sekumpulan sumber data yang terdiri dari *database*, *data warehouse*, *world wide web* atau media penyimpanan lain, seperti *spreadsheet*, dan lain-lain.

Sumber data ini kemudian diolah melalui serangkaian proses *data cleaning*, proses *data integration*, dan proses pemilahan. Proses ini akan dibahas pada sub bab berikutnya.

Pada lapisan kedua, terdiri dari *database server* atau *data warehouse server* yang bertanggung jawab untuk mengambil data yang relevan, berdasarkan kebutuhan pengguna yang merupakan hasil dari proses di atas.

Knowledge base adalah kumpulan bidang pengetahuan yang dipergunakan untuk dijadikan acuan untuk mencari atau mengevaluasi kemenarikan dari suatu pola yang dihasilkan. Beberapa pengetahuan melibatkan hirarki konsep, yang digunakan untuk mengorganisasi atribut-atribut atau nilai-nilai atribut tersebut ke beberapa level abstraksi yang berbeda. Pengetahuan lain seperti kepercayaan pengguna, yang dapat digunakan untuk menilai kemenarikan suatu pola yang tak terduga, juga bisa dilibatkan. Contoh lain dari bidang pengetahuan adalah kemenarikan batasan atau ambang batas dan *metadata* (contoh metadata: data yang menjelaskan dari mana suatu data diambil).

Data mining engine adalah hal paling penting dan secara ideal terdiri dari kumpulan modul fungsional untuk beberapa pekerjaan seperti karakterisasi, asosiasi, dan analisis korelasi, klasifikasi, prediksi, analisis kluster, analisis *outlier*, dan analisis evolusi.

Pattern evaluation module adalah modul yang menerapkan pengukuran terhadap suatu kemenarikan pola dan berinteraksi dengan modul-modul pada *data mining engine* yang dapat mencari pola yang menarik. Ambang batas kemenarikan dapat digunakan untuk menyaring pola yang diketemukan.

User interface adalah modul yang berkomunikasi antara pengguna dengan system *data mining*, dimana pengguna dapat menentukan *query*, menyediakan informasi pencarian dan melakukan eksplorasi *data mining*. Sebagai tambahan pengguna dapat mempelajari pola dan memvisualkan pola dalam beberapa bentuk.

2.3 Teknik Klasifikasi

Teknik klasifikasi bisa disimpulkan sebagai cara memprediksi suatu data baru sehingga bisa ditentukan pada kategori apakah ia berada, berdasarkan pada data latih, dimana tiap anggota data latih tersebut telah diketahui kategorinya. Kategori ini tentunya bersifat diskrit, dimana urutan tidak mempengaruhi [9]. Contohnya seperti: positif, negatif, dan netral; baik dan buruk; dll.

2.4 Proses Klasifikasi

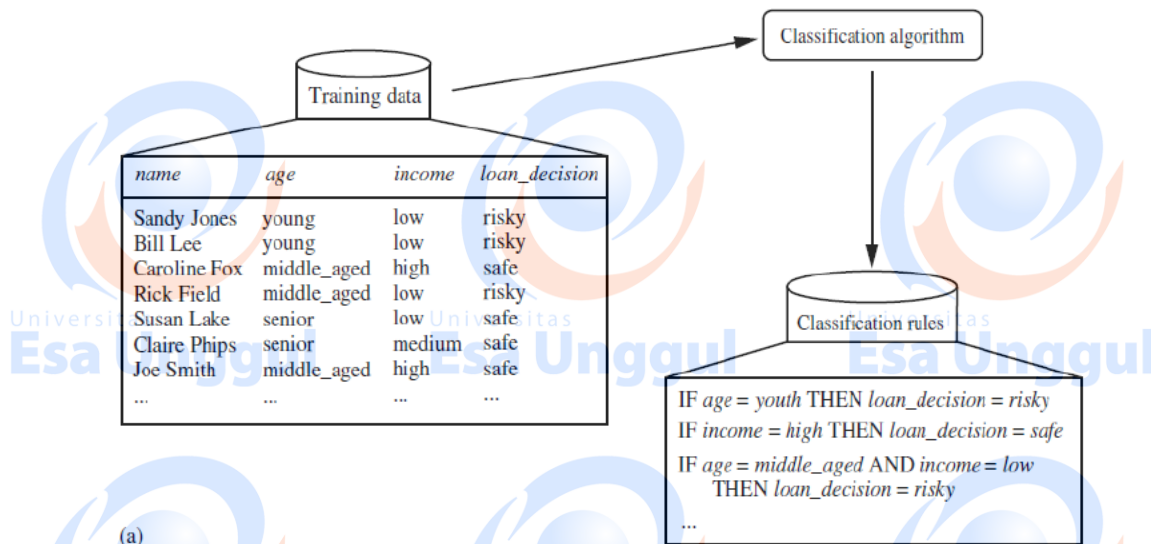
Dalam teknik klasifikasi ada dua proses utama yaitu proses pembangunan model dan penerapan model [9]. Proses pembangunan model melibatkan tahapan sebagai berikut:

1. Menentukan kategori/kelas/label terlebih dahulu. Misal: positif, negatif, dan netral.
2. Dari sekumpulan data yang diperoleh, tentukan kategori untuk tiap datanya.
3. Sekumpulan data yang telah dikategorisasikan ini disebut dengan data latih yang akan digunakan sebagai model.
4. Model ini bisa digambarkan sebagai aturan klasifikasi, pohon keputusan atau formula matematika.
5. Algoritma berdasarkan model di atas untuk mengklasifikasi disebut dengan *classifier* (pengklasifikasi).

Proses ini dapat disebut juga sebagai *supervised learning* (pelatihan terawasi). Disebut terawasi karena tiap datanya sudah diberikan label. Proses yang kedua adalah proses penerapan model atau proses klasifikasi. Proses ini melibatkan tahap:

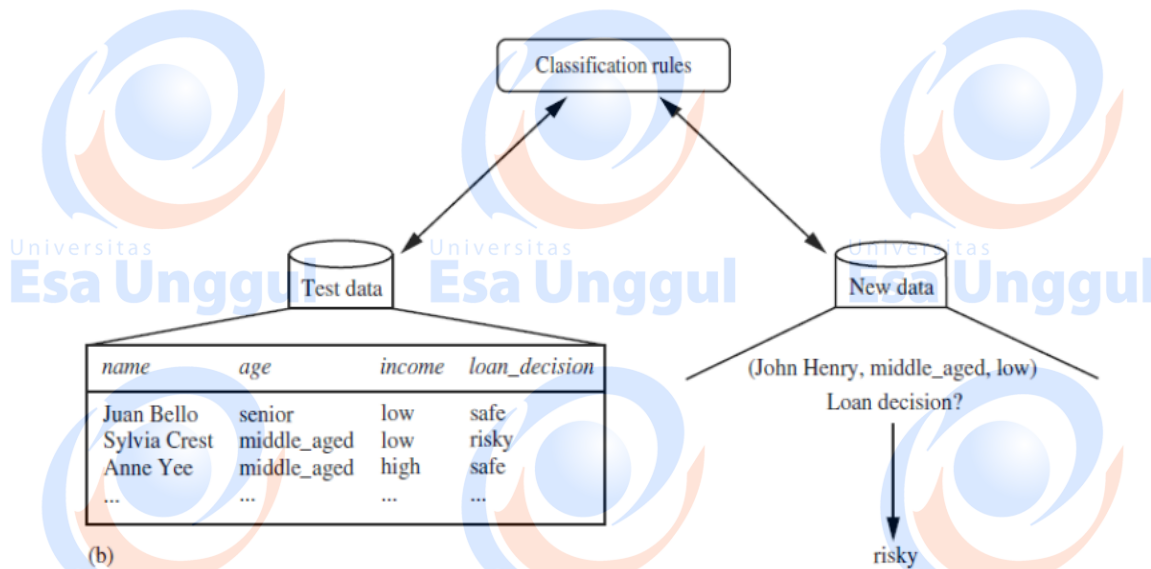
1. Tentukan sekumpulan data untuk diuji.
2. Sekumpulan data uji ini tiap datanya telah diberikan kategori/kelas/label.
3. Dilakukan proses pemetaan dengan menggunakan *classifier* di atas. Data uji ini akan ditentukan kategorinya berdasarkan model di atas dan kemudian hasilnya dibandingkan dengan kategori yang telah diberikan. Misal: pada data uji, dinyatakan bahwa data X adalah positif. Setelah dilakukan proses klasifikasi dengan menggunakan data latih ternyata data X bernilai negatif.
4. Kemudian ditentukan akurasi model di atas dengan menghitung seberapa banyak kategori yang dihasilkan bernilai sama dengan kategori yang telah ditentukan pada data uji di awal.
5. Jika rasio akurasi memuaskan (memenuhi batas minimal yang ditentukan), maka *classifier* tersebut dapat digunakan untuk data baru.

Untuk lebih jelasnya gambar di bawah ini bisa menjelaskan proses tersebut di atas.



Gambar 2.2. Proses Pembangunan Model [9]

Training data atau data latih, dengan algoritma klasifikasi dihasilkan *classification rules* atau aturan klasifikasi yang disebut dengan *classifier* (pengklasifikasi). Pada contoh data di atas, kolom *loan_decision* adalah label atau kategori yang telah ditentukan.

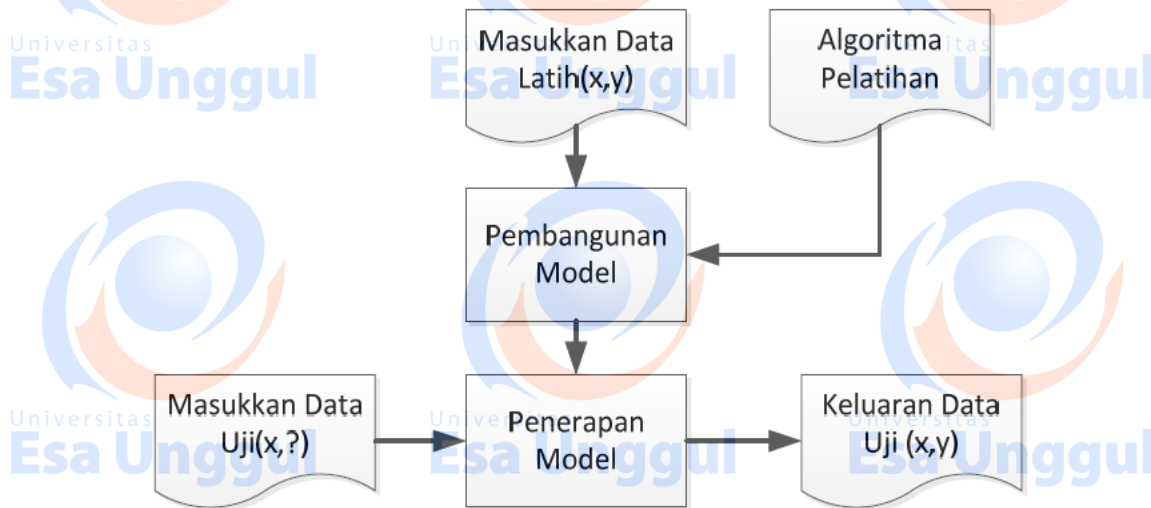


Gambar 2.3. Proses Penerapan Model [9]

Dengan *classifier* yang telah dihasilkan, diterapkan pada *test data* atau data uji untuk diukur keakuratannya. Hasil akurasi adalah perbandingan dari jumlah total hasil klasifikasi menggunakan classifier pada data uji yang nilainya sama dengan nilai kolom *loan_decision* pada data uji. Jadi misal untuk data (*Juan Bello, senior, low*), jika diproses dengan *classifier* akan menghasilkan nilai

label/kategori *safe*. Hal ini sama dengan nilai aslinya. Karena sama, hasil itu ikut dihitung akurasi. Jika hasilnya tidak sama, maka tidak dihitung. Jumlah total akurasi dibagi jumlah data uji menjadi hasil akurasi. Jika hasil akurasi dapat memenuhi ambang batas yang telah ditetapkan, maka *classifier* siap diterapkan pada data baru.

Secara lebih ringkas, proses tersebut bisa digambarkan sebagai berikut:



Gambar 2.4. Proses Pekerjaan Klasifikasi [10]

2.5 Algoritma Klasifikasi

Dalam membantu pekerjaan klasifikasi ada beberapa algoritma klasifikasi yang telah disusun oleh beberapa pakar peneliti. Liu [8] menyatakan analisis sentimen adalah mengklasifikasikan text, maka algoritma yang paling cocok adalah algoritma *Naïve Bayes* dan *Support Vector Machine (SVM)*. Algoritma *Naïve Bayes* merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema *Bayes* dengan asumsi independensi yang kuat atau naif [10]). Sedangkan algoritma SVM merupakan teknik hasilnya lebih menjanjikan dan memberikan metode yang lebih baik dari yang lain namun lebih rumit. Penulis memutuskan untuk menggunakan algoritma *Naïve Bayes* dengan pertimbangan kesederhanaan dan kemudahan dalam penerapannya.

Algoritma Naïve Bayes

Teorema *Bayes* mempunyai formula umum sebagai berikut:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

dimana:

1. $P(H/E)$ adalah probabilitas akhir bersyarat (*posterior probability*) suatu hipotesis H terjadi pada jika diberikan bukti E terjadi.
2. $P(E/H)$ adalah probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H.
3. $P(H)$ adalah probabilitas awal (*prior probability*) hipotesis H terjadi tanpa memandang bukti apapun.
4. $P(E)$ adalah probabilitas awal (*prior probability*) bukti E tanpa memandang hipotesis/bukti yang lain.

Sebagaimana telah dijelaskan sebelumnya, bahwa sentimen/opini terdiri dari entiti target, aspek/fitur, nilai sentimen, pemilik sentimen, dan waktu sentimen dibuat. Untuk menggunakan teori *Bayes*, dua variabel yang dipakai adalah aspek/fitur sebagai hipotesis (H) dan nilai sentimen sebagai bukti (E). Tiga variabel lainnya akan digunakan sebagai metadata dari sentimen tersebut.

Karena dalam suatu kalimat terdiri dari banyak kata, dimana sangat sulit dalam praktiknya untuk menentukan mana yang bisa disebut sebagai aspek/fitur, maka diasumsikan bahwa setiap kata adalah aspek/fitur.

Maka penerapan teori *Bayes* adalah sebagai berikut:

$$P(K|F) = \frac{P(F|K) \times P(K)}{P(F)}$$

dimana:

1. F adalah fitur atau kata.
2. K adalah kategori atau nilai sentimen.

Karena fitur yang mendukung satu kategori bisa banyak, misal ada fitur F_1, F_2, F_3 , maka teori *Bayes* dapat dikembangkan menjadi:

$$P(K|F_1, F_2, F_3) = \frac{P(F_1, F_2, F_3|K) \times P(K)}{P(F_1, F_2, F_3)}$$

Karena teori *Naïve Bayes* mensyaratkan bahwa bukti-bukti (dalam hal ini fitur-fitur) yang ada adalah independen satu sama lain maka bentuk rumus di atas bisa diubah menjadi:

$$P(K|F_1, F_2, F_3) = \frac{P(F_1|K) \times P(F_2|K) \times P(F_3|K) \times P(K)}{P(F_1) \times P(F_2) \times P(F_3)}$$

3. TAHAPAN PENELITIAN

Penelitian ini memiliki tahapan sebagai berikut:

- Tahap Pengumpulan Data
Pada tahap ini, dilakukan pengumpulan data yang diperoleh dari media sosial Twitter dengan menggunakan aplikasi R Studio.
- Tahap Praproses Data
Tahap praproses data dimana data yang sudah dikumpulkan akan dilakukan pembersihan data dengan cara *remove duplicate*, *replace/filtering*, *transform cases/case folding* dan *tokenizing*.
- Tahap Klasifikasi Data
Pada tahap ini, data yang sudah dibersihkan melalui tahap praproses data akan diklasifikasikan berdasarkan sentimen dari data yang telah dikumpulkan. Adapun sentimen yang digunakan untuk klasifikasi yaitu sentimen positif, netral, dan negatif.
- Tahap *Naive Bayes Classifier*
Pada Tahap ini, data yang sudah diklasifikasikan akan di proses menggunakan metode *naive bayes classifier*, dimana akan diketahui tingkat akurasi terhadap data tersebut.
- Tahap Kesimpulan
Setelah data training diklasifikasikan selanjutnya akan dibandingkan dengan data KPU. Hasilnya akan ditarik kesimpulan apakah ada korelasi antara data sentimen dari twitter dengan data KPU.

Secara lebih rinci, tahapan penelitian ini digambarkan pada Gambar 3.1.

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

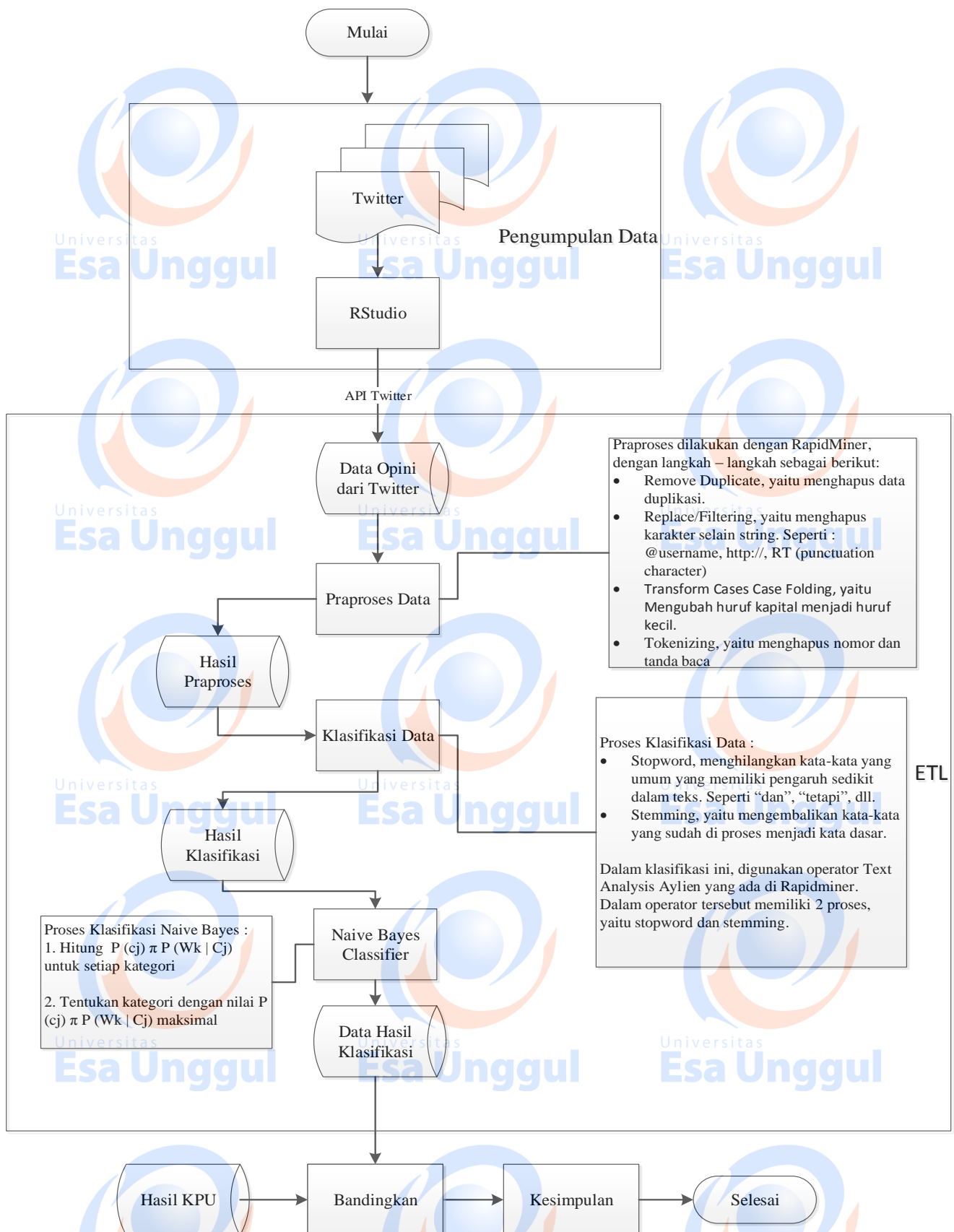
Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul

Universitas
Esa Unggul



Gambar 3.1. Tahapan Penelitian

4. Hasil dan Pembahasan

4.1. Pengumpulan Data Twitter

Proses *crawling* / pengumpulan data dilakukan dengan menggunakan *web crawler* R Studio dengan menggunakan kata kunci yang sudah ditentukan. Hasil dari proses *crawling* data twitter dari 20 Juni – 20 Juli 2018 didapatkan data *tweet* sebanyak 11.527.

4.2. Pra Proses Data

Pra-proses data dilakukan sebelum proses klasifikasi supaya dimensi *vector space model* menjadi lebih rendah. Dengan membuat dimensi *vector space model* menjadi lebih rendah proses klasifikasi akan menjadi lebih cepat.

Untuk mendapatkan data yang bersih yang bisa di klasifikasi, dilakukan hal-hal berikut:

- *Remove Duplicate*, yaitu menghapus data duplikasi / berulang.
- *Replace / Filtering*, berupa penghapusan semua karakter selain string serta penghapusan beberapa karakteristik dari data, misalnya @username, #hashtag, http://, dan “RT”. Dalam *filtering* juga dilakukan penghapusan terhadap *stopward*. Hal ini bermanfaat untuk mengurangi *load* dan *performance* saat melakukan training maupun testing data.
- *Transform Cases*, yaitu mengubah semua huruf kapital menjadi huruf kecil atau *lowercase*.
- *Tokenizing*, yaitu pemecahan berdasarkan perkata. Hal ini bisa dilakukan dengan menandai karakter sebagai pembatas. Tahapan yang dilakukan adalah menghapus nomor dan tanda baca.

Hasil yang didapat setelah dilakukan pra proses bisa dilihat pada Tabel 1 berikut:

Tabel 4.1. Hasil *Crawl* data Twitter

No	Keyword	Data Mentah	Data Praproses
1	Pilkadajabar2018	1.063	24
2	Pilgubjabar2018	3.638	12
3	Ridwan-uu	1.158	20
4	Hasanah	541	42
5	Asyik	4.253	175
6	Deddy-dedi	825	23
Total Data		11.527	296

4.3. Proses Klasifikasi

Data hasil pra proses yang sudah bersih selanjutnya diklasifikasikan dengan Rapid Miner dengan teknik *naive bayes classifier*. Berikut ini adalah hasil klasifikasi dengan menggunakan Rapid Miner

a. Pasangan Rindu

accuracy: 95.00% +/- 15.00% (micro average: 94.44%)

	true negative	true positive	class precision
pred. negative	3	0	100.00%
pred. positive	1	14	93.33%
class recall	75.00%	100.00%	

Gambar Error! No text of specified style in document..1. Tingkat Akurasi Pasangan Rindu

b. Pasangan Hasanah

accuracy: 84.17% +/- 17.26% (micro average: 84.21%)

	true negative	true positive	class precision
pred. negative	0	3	0.00%
pred. positive	3	32	91.43%
class recall	0.00%	91.43%	

Gambar Error! No text of specified style in document..2 Tingkat Akurasi Pasangan Hasanah

c. Pasangan Asyik

accuracy: 79.08% +/- 9.65% (micro average: 79.08%)

	true negative	true positive	class precision
pred. negative	8	19	29.63%
pred. positive	13	113	89.68%
class recall	38.10%	85.61%	

Gambar Error! No text of specified style in document..3. Tingkat Akurasi Pasangan Asyik

d. Pasangan Deddy-Dedi

accuracy: 90.00% +/- 20.00% (micro average: 90.00%)

	true negative	true positive	class precision
pred. negative	0	1	0.00%
pred. positive	1	18	94.74%
class recall	0.00%	94.74%	

Gambar Error! No text of specified style in document..4. Tingkat Akurasi Pasangan Deddy-Dedi

Secara ringkas hasil dari pengolahan klasifikasi dengan Rapid Miner didapatkan hasil sebagai berikut:

Tabel 4.2. Hasil Klasifikasi Data Twitter

No	Keyword	Tingkat Akurasi
1	Pilkadajabar2018	90.00%
2	Pilgubjabar2018	95.00%
3	Rindu	95.00%
4	Hasanah	84.17%
5	Asyik	79.08%
6	Deddy-Dedi	90.00%

4.4. Pembahasan

Merujuk pada penelitian Simada, et all [13], klasifikasi sentimen menggunakan *naive bayes classifier* tingkat akurasi dapat digunakan untuk mengukur *preference value* pada kasus pemilihan kepala daerah sehingga mendapatkan respon positif untuk pasangan Gubernur dan Wakil Gubernur Jawa Barat 2018. Untuk menghindari bias, jumlah sentimen netral tidak dilibatkan dalam perhitungan. Berdasarkan hal tersebut didapatkan lah hasil seperti pada tabel 4.3

Tabel 4.3. Perbandingan Data Sentimen

No	Keyword	Positif		Negatif		Total Data
1	Rindu	14	77.77%	4	22.23%	18
2	Hasanah	35	92.10%	3	7.90%	38
3	Asyik	132	85.71%	22	14.29%	154
4	Deddy-Dedi	19	95.00%	1	5.00%	20

Selanjutnya, data sentimen yang positif tersebut dibandingkan dengan data perolehan suara dari KPU guna melihat ada tidaknya korelasi antara analisis sentimen dengan perolehan suara di pilkada, dimana hasilnya seperti terlihat pada Tabel 4.4.

Tabel 4.4. Tabel Perbandingan Antara Data Analisis Sentimen dengan Perolehan Suara KPU

No	Keyword	Data Twitter		Perolehan Suara KPU
		Akurasi	Positif	
1	Rindu	95.00 %	77.77%	32.88 %
2	Hasanah	84.17 %	92.10%	12.62 %
3	Asyik	74.08 %	85.71%	28.74 %
4	Deddy-Dedi	90.00 %	95.00%	25.77 %

Dari Tabel 4.4 terlihat bahwa pasangan Rindu mendapatkan respon positif tertinggi sedangkan pasangan Asyik mendapat respon positif terendah. Hal ini berbeda dengan hasil perolehan suara KPU dimana pasangan Rindu mendapatkan suara terbanyak, namun pasangan Asyik justru mendapat perolehan suara nomor dua. Hal ini menunjukkan bahwa tidak ada korelasi antara data analisis sentiment data twitter dengan perolehan suara KPU.

Tidak adanya korelasi ini bisa jadi disebabkan oleh data riil yang ada di twitter dimana

- Data opini Twitter banyak yang isinya sama
- Banyaknya data hasil RT (ReTweet) dari user lain
- Banyaknya waktu *tweet* yang hampir sama
- Banyaknya user yang tidak lazim.

Kondisi tersebut sangat mempengaruhi hasil pra-proses, dimana akan mengakibatkan banyaknya data yang dihilangkan. Dampaknya hanya sedikit data bersih yang bisa diproses untuk analisis sentimen. Hasilnya tidak ada korelasi antara data analisis sentimen dengan data perolehan suara KPU. Berikut adalah contoh hasil perolehan data mentah dari twitter yang menggambarkan kondisi tersebut.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
301	01/07/2018 12.57	1,01E+35 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
302	01/07/2018 12.56	1,01E+34 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
303	01/07/2018 12.53	1,01E+35 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
304	01/07/2018 12.49	1,01E+35 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
305	01/07/2018 12.48	1,01E+35 #Asyik di Jabar blm terima hsl QC krm byk kejanggalan, tanpa aksi bakar2, nunggu hsl KPU, Jokowi nyinyir sampai d... https://t.co/dykyG69Jlq																		
306	01/07/2018 12.47	1,01E+35 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
307	01/07/2018 12.45	1,01E+35 Aksesoris kalung cantik wanita. https://t.co/VN884Lo8n #PilisJanganDicopas #wanitacantik #wanita #buhamil #jual... https://t.co/6900N7F9T																		
308	01/07/2018 12.45	1,01E+35 RT @ButirHikmah: Sebelum pilkada @ridwanKamil sebar pernyataan yg membantah dukungan pdp kju utk menarik hati pemilih Jabar. Setelah menang...																		
309	01/07/2018 12.42	1,01E+35 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
310	01/07/2018 12.42	1,01E+35 RT @BungSalingSapa: Demi Allah, tdk ada kader yg berkata 'saya mau dapat posisi apabila #Asyik menang' namun mereka selalu berkata 'afwan ak...																		
311	01/07/2018 12.42	1,01E+35 RT @BungSalingSapa: Tiap pekan para kader mengadakan serangan dg berbagai media dg biaya patungan, ada pula yg menyumbang nasi bungkus, air...																		
312	01/07/2018 12.41	1,01E+35 RT @BungSalingSapa: Salah satu tanda kemenangan partai Dakwah ini adalah kembali kokohnya ukhuwah, militansi kader di tengah upaya adu domb...																		
313	01/07/2018 12.40	1,01E+34 RT @BungSalingSapa: Demi Allah, tdk ada kader yg berkata 'saya mau dapat posisi apabila #Asyik menang' namun mereka selalu berkata 'afwan ak...																		
314	01/07/2018 12.38	1,01E+35 RT @BungSalingSapa: Salah satu tanda kemenangan lain adalah logika logika & nilai dakwah PKS sdh dipahami & dukung rakyat Indonesia khusus...																		
315	01/07/2018 12.37	1,01E+35 RT @BungSalingSapa: Tiap pekan para kader mengadakan serangan dg berbagai media dg biaya patungan, ada pula yg menyumbang nasi bungkus, air...																		
316	01/07/2018 12.36	1,01E+35 RT @BungSalingSapa: Demi Allah, tdk ada kader yg berkata 'saya mau dapat posisi apabila #Asyik menang' namun mereka selalu berkata 'afwan ak...																		
317	01/07/2018 12.35	1,01E+34 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
318	01/07/2018 12.35	1,01E+35 RT @BungSalingSapa: Siapapun boleh nyinyir, namun kami para kader tak gentar sedikitpun, kami senantiasa #Asyik berjuang dlm dakwah ini hin...																		
319	01/07/2018 12.35	1,01E+35 RT @BungSalingSapa: Salah satu tanda kemenangan lain adalah logika logika & nilai dakwah PKS sdh dipahami & dukung rakyat Indonesia khusus...																		
320	01/07/2018 12.34	1,01E+35 RT @BungSalingSapa: Salah satu tanda kemenangan partai Dakwah ini adalah kembali kokohnya ukhuwah, militansi kader di tengah upaya adu domb...																		
321	01/07/2018 12.34	1,01E+35 RT @BungSalingSapa: Kamsu kami para kader, tdk ada yg kalah dlm memperjuangkan partai dakwah, justru kekalahan jika kami berkhianat dlm ber...																		
322	01/07/2018 12.34	1,01E+35 RT @BungSalingSapa: Demi Allah, tdk ada kader yg berkata 'saya mau dapat posisi apabila #Asyik menang' namun mereka selalu berkata 'afwan ak...																		
323	01/07/2018 12.34	1,01E+35 RT @BungSalingSapa: Tiap pekan para kader mengadakan serangan dg berbagai media dg biaya patungan, ada pula yg menyumbang nasi bungkus, air...																		
324	01/07/2018 12.34	1,01E+34 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
325	01/07/2018 12.33	1,01E+35 RT @BungSalingSapa: Demi Allah, Bukan kekalahan yg membuat kami menangis, tapi kami menangis menyaksikan semangat juang kader tuk memenang...																		
326	01/07/2018 12.33	1,01E+35 RT @Aiat: Syaikh: Pesan Ustadz @syaikh_ahmad kepada para kader, relawan dan simpatisan #Asyik https://t.co/11Hfio5tY																		

Gambar 4.5. Contoh Data Mentah Hasil Crawling Data Twitter

5. Kesimpulan dan Saran

5.1. Kesimpulan

Dari hasil pembahasan sebelumnya bisa diambil beberapa kesimpulan sebagai berikut:

1. Jumlah data tweet tentang pilkada Jawa Barat yang berhasil dikumpulkan dari tanggal 20 Juni sampai dengan 20 Juli 2018 didapatkan data 11.527 dengan 6 *keywords*. Hasil dari pra proses atas data tersebut didapatkan 296 data bersih yang bisa diklasifikasikan
2. Dengan menggunakan Rapid Miner dan operator analisis sentimen dari Aylie bisa didapatkan data analisis sentimen twitter dimana akurasi tertinggi diperoleh pasangan Rindu (95.00 %), sedangkan hasil dengan akurasi terendah diperoleh pasangan Asyik (74.08%). Hal ini berbeda dengan data perolehan suara KPU dimana perolehan suara tertinggi adalah pasangan Rindu (32.88%) dan pasangan Asyik mendapatkan posisi kedua (28.74%). Dengan demikian tidak ada korelasi antara analisis sentimen data twitter dengan perolehan suara KPU dalam pilkada
3. Tidak adanya korelasi ini bisa jadi disebabkan oleh banyaknya duplikasi tweet, banyak akun abal-abal serta banyaknya ReTweet (RT) sehingga saat dilakukan pra proses, jumlah data yang ada menjadi berkurang secara signifikan.

5.2. Saran

Untuk kesempurnaan penelitian ini, berikut ini adalah saran perbaikan guna mendapatkan hasil yang lebih komprehensif

1. Untuk mendeteksi data tweet yang berulang, akun abal-abal dan berbagai kondisi seperti yang sudah disebutkan diatas, alangkah baiknya jika digunakan pendekatan lain yaitu SNA (Social Network Analyzer)
2. Agar pra proses bisa menghasilkan data yang lebih bersih lagi, perlu adanya analisis hubungan antar kata yang sesuai dengan KBBI. Dengan demikian hasil pra proses bisa sesuai dengan kaidah Bahasa Indonesia baku

DAFTAR PUSTAKA

1. Cahyanti, O.D., Saksono, P.H., Suryayusra, Negara, E.S., (2015). Social Media Analytics Pemanfaatan Data Media Sosial Untuk Penelitian, Palembang.
2. Grafelly, Delvit Bagaimana perkembangan Twitter saat ini?. Diakses 10 Desember 2015, dari <http://www.techno.id/social/bagaimana-perkembangan-twitter-saat-ini-1509122.html>.
3. Rozi, IF., Pramono, S.H., dan Dahlan, E. A. (2012). Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. Jurnal EECCIS Vol. 6, No. 1, Juni 2012.
4. Kamaruzaman, S.M., Chowdhury M.R. 2004. *Text Categorization using Association Rule and Naive Bayes Classifier*. Asian Journal of Information Technology, Vol. 3, No. 9, pp 657-665, Sep. 2004
5. Kibriya Ashraf M., Frank Eibe, Pfahringer Bernhard. Holmes Geoffrey . 2004. *Multinomial Naive Bayes for TextCategorization Revisited*. Australian joint conference on artificial intelligence No 17.
6. Femphy Pisceldo, Manurung, R., Adriani, Mirna. 2009. *Probabilistic Part-of-Speech Tagging for bahasa Indonesia*. Third International MALINDO Workshop, colocated event ACLIJCNLP 2009, Singapore, August 1, 2009.
7. Wicaksono, Alfani F dan Purwarianti, Ayu. 2010. *HMM Based Part-of-Speech Tagger for Bahasa Indonesia*. Proceeding of the Fourth International MALINDO Workshop (MALINDO2010). Agustus 2010. Jakarta, Indonesia
8. Bing Liu. *Sentiment analysis and opinion mining*. (2012). Morgan & Claypool Publishers.
9. Jiawei Han and Micheline Kamber. (2006). *Data mining: concepts and techniques*. Second Edition, San Francisco: Morgan Kaufmann
10. Eko Prasetyo. *Data Mining-Konsep dan Aplikasi menggunakan Matlab*. (2012). Edisi ke-1, Yogyakarta: ANDI
11. Hemalatha, I., Varma, P.G., dan Govardhan, A. (2012). Preprocessing the Informal Text for Efficient Sentiment Analysis, *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, Vol.1, July – August 2012, ISSN 2278-6856
12. Triawati, Chandra. (2009). *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia*, Institut Teknologi Telkom, Bandung.
13. Simada H., Ginting, Lhaksmana., Kemas Muslim & Murdiansyah, Danang Triantoro (2018). *Klasifikasi Sentimen Terhadap Bakal Calon Gubernur Jawa Barat 2018 di Twitter Menggunakan Naive Bayes*. E-proceeding of engineering : Vol.5, No 1. 1793.

