Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier

Faishol Nurhuda
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No 36 A, Surakarta
f@ishol.net

Sari Widya Sihwi
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No 36 A, Surakarta
sari.widya.sihwi@gmail.com

Afrizal Doewes
Jurusan Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No 36 A, Surakarta
afrizal.doewes@staff.uns.ac.id

ABSTRAK

Dalam penelitan ini akan dilakukan analisis sentimen masyarakat terhadap calon presiden dan wakil presiden Indonesia 2014 yang diungkapkan melalui jejaring sosial Twitter. Ada beberapa tahap untuk melakukan analisis sentimen, diantaranya adalah tahap pengumpulan data, praprocesing data, POS Tagging, ekstrasi opini menggunakan rule based dan klasifikasi opini menggunakan metode Naïve Bayes Classifier.

Hasil dari penelitian ini didapatkan bahwa pasangan capres dan cawapres Prabowo Subianto dan Hatta Rajasa mendapatkan jumlah percakapan 53% dan pasangan Joko Widodo – Jusuf Kalla mendapatkan 47%. Sedangkan untuk hasil polaritas sentimen, Prabowo Subianto – Hatta Rajasa mendapatkan 47,7% untuk sentimen positif, 26,4% sentimen negatif dan 25,9% sentimen netral. Sedangkan pasangan Joko Widodo – Jusuf Kalla mendapatkan total 37,6% sentimen positif, 34,4% sentimen negatif, dan 27,9 sentimen netral.

Kata kunci: analisis sentimen, pemilu, *naïve bayes classifier*

1. PENDAHULUAN

Indonesia adalah salah satu negara yang menganut sistem demokrasi. Hal ini ditandai dengan diadakannya suatu pemilihan umum terhadap presiden dan wakil presiden. Pemilihan umum pada suatu negara yang menganut demokrasi bisaanya diselenggarakan secara periodik. Pada tahun 2014 akan dilaksanakan pemilihan umum presiden dan wakil presiden. Seorang tokoh politik yang ingin maju sebagai calon presiden tentu akan melihat atau mempertimbangkan popularitas mereka berdasarkan opini dari masyarakat. Dahulu masyarakat mengungkapkan opini, kritik, dan sarannya melalui media cetak yang tidak semua orang mempunyai kemampuan menulis dan kesempatan menerbitkan tulisannya. Namun, perkembangan teknologi komunikasi saat ini telah merubah kecenderungan kebisaaan masyarakat dalam mengekspresikan opininya pada jejaring sosial. Salah satu jejaring sosial yang populer di kalangan pengguna internet saat ini adalah Twitter.

Perkembangan penggunaan Twitter sangat cepat. Hal ini berdasarkan hasil survey *PeerReach* yang dilansir situs Beritagar.com[1] pada 15 November 2013, yaitu terdapat lebih dari 100 juta pengguna aktif di seluruh dunia bersama-sama mengirimkan 250 juta tweet setiap hari. Pengguna Twitter berasal dari Indonesia hingga Oktober 2013 menjadi yang terbanyak ketiga di dunia. Kesederhanaan dan kemudahan dalam penggunaan merupakan beberapa alasan mengapa Twitter lebih digemari masyarakat Indonesia dalam berkomunikasi. Setiap pengguna Twitter bebas mem-*post* tweet dengan batasan 140 karakter.

Tweet adalah teks status pengguna yang digunakan untuk memberikan informasi di Twitter. Berdasarkan kutipan hasil penelitian[2], tweet bisaa digunakan penggunanya untuk memberitahu tentang apa yang sedang dilakukan atau dirasakan, percakapan, berbagi informasi, dan pelaporan berita. Pada umumnya tweet digunakan untuk mem-post hal tentang diri pengguna dan berbagi informasi. Isi tweet juga dapat mengekspresikan perasaan atau *mood* pengguna, misalkan "Aku sangat suka dengan pribadi beliau", hal ini bersifat penilaian subjektif atau opini. Opini melalui tweet inilah yang dapat dimanfaatkan untuk melihat bagaimana sentimen yang dimunculkan salah satunya adalah mengenai opini seseorang terhadap tokoh politik yang akan maju sebagai calon presiden Indonesia tahun 2014.

Penentukan polaritas positif atau negatifnya suatu opini dapat dilakukan secara manual, tetapi seiring bertambahnya sumber opini menjadi semakin banyak tentunya waktu dan usaha yang dibutuhkan untuk mengklasifikasikan polaritas opini tersebut akan semakin banyak terpakai. Oleh karena itu, diajukan penerapan metode pembelajaran mesin untuk mengklasifikasi polaritas opini dari sumber data yang sangat banyak tersebut. Untuk melakukan hal itu, bisa menggunakan salah satu fungsi dari *text mining*, dalam hal ini adalah klasifikasi dokumen.

Ada beragam teknik klasifikasi dokumen, di antaranya adalah *Naïve Bayes classifier*, *Decision Trees*, dan *Support Vector Machines*. Salah satu metode yang paling populer digunakan dalam pengklasifikasian dokumen sekarang ini adalah metode *Naïve Bayes classifier*[3]. *Metode Naïve Bayes classifier* mempunyai kecepatan dan akurasi yang tinggi ketika diaplikasikan dalam basis data yang besar dan data yang beragam[4]. Hal serupa juga diungkapkan oleh [5] dalam penelitiannya, yaitu *metode Naïve Bayes Classifier* memiliki beberapa kelebihan antara lain, sederhana, cepat dan berakurasi tinggi.

Penelitian mengenai analisis sentimen telah banyak dilakukan sebelumnya. Diantaranya adalah yang melakukan klasifikasi sentimen terhadap *review* film dengan menggunakan berbagai teknik pembelajaran mesin. Teknik pembelajaran mesin yang digunakan yaitu *Naïve Bayes, Maximum Entropy, dan Support Vector Machines* (SVM). Penelitian tetang analisis sentimen yang menggunakan dataset dari jejaring sosial Twitter dilakukan oleh [6]. Mereka melakukan analisis

sentimen terhadap media jejaring sosial Twitter dengan menggunakan beberapa teknik klasifikasi. Penelitian selanjutnya yang menjadi acuan penulis dalam menyusun penelitian ini adalah penelitian yang dilakukan oleh [7]. Pada penelitian tersebut dilakukan analisis sentimen menggunakan metode *naïve bayes* dalam penentuan polaritas sentimen. Hasil dari peneltian tersebut menunjukkan akurasi yang cukup tinggi untuk metode *Naïve Bayes Classifier.* Atas dasar hal ini, penulis bermaksud menerapkan metode *Naïve Nayes Classifier* untuk melihat sentimen masyarakat di media Twitter terhadap tokoh politik yang maju dalam pemilihan umum presiden dan wakil presiden Indonesia tahun 2014.

2. DASAR TEORI

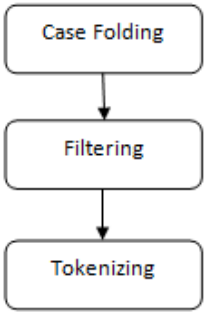
2.1. Text Mining

Text mining juga dikenal sebagai *data mining* teks [8] atau penemuan pengetahuan dari database tekstual [9]. Sesuai dengan buku *The Text Mining Handbook* [10], text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam data mining. Tujuan dari *text mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan *dalam text mining* adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks dan pengelompokkan teks [11].

Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian / pengelompokkan dan menganalisa *unstructured data* dalam jumlah besar, dalam hal ini data yang akan digunakan adalah data yang diambil dari Twitter. Dalam memberikan solusi, *text mining* mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *Data Mining*, *Information Retrieval*, Statistik dan Matematik, *Machine Learning, Linguistic*, *Natural Language Processing dan Visualization*. Kegiatan riset untuk *text mining* antara lain ekstraksi dan penyimpanan teks, *preprocessing* akan konten teks, pengumpulan data statistik serta *indexing* dan analisis sentimen[11].

2.2. Ekstraksi Dokumen

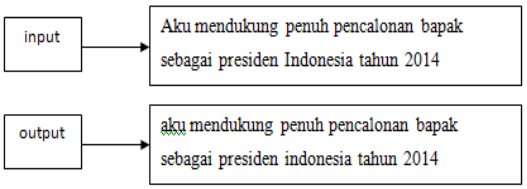
Teks yang akan dilakukan proses *text mining*, pada umumnya memiliki karakteristik diantaranya memiliki dimensi yang tinggi, terdapat *noise* pada data, dan terdapat struktur teks yang tidak baik. Dalam hal ini yang digunakan data yang berasal dari Twitter. Data yang bersasal dari Twitter mempunyai kerumitan yang cukup tinggi. Hal ini di karenakan karakteristik dari Twitter adalah penggunaan bahasa yang tidak sesuai bahasa baku dan banyaknya kesalahan ejaan pada penulisan *tweet*[12]. Cara yang digunakan dalam
mempelajari suatu data teks, adalah dengan terlebih dahulu menetukan fitur-fitur yang mewakili setiap kata untuk setiap fitur yang ada pada dokumen. Menurut [13], sebelum menentukan fitur-fitur yang mewakili, diperlukan tahap *preprocessing* yang dilakukan secara umum dalam *text mining* pada dokumen, yaitu *case folding, tokenizing,* dan *filtering* seperti yang ditunjukkan pada gambar 2.1.



Gambar 2.1 Proses Ekstraksi Dokumen

2.3. Case Folding

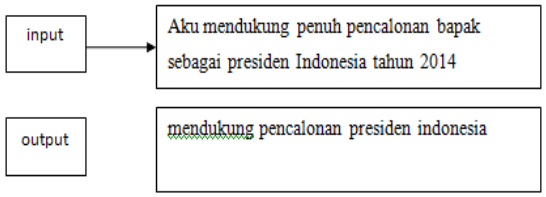
Case folding adalah mengubah semua huruf besar atau kapital dalam Twitter menjadi huruf kecil[12]. Hanya huruf 'a' sampai 'z' yang diterima. Karakter selain huruf dihilangkan dan dianggap *delimiter*. *Delimiter* adalah urutan satu karakter atau lebih yang dipakai untuk membatasi atau memisahkan data yang disajikan dalam *plain text* Contoh dari tahap ini seperti yang ada dalam Gambar 2.2.



Gambar 2.2. Proses *Case Folding*

2.4. Filtering

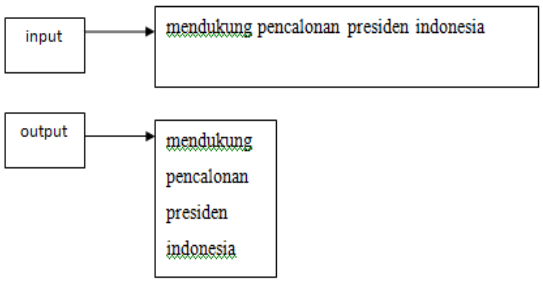
Tahap *filtering* adalah tahap mengambil kata-kata penting dari hasil token. Pada tahap ini akan dilakukan pembersihan *tweet* dari spesial karakter, *URL link*, username, serta *emoticon*. Berdasarkan penelitian dari[12], *emoticon* disini dihilangkan juga karena akan mempengaruhi hasil akurasi yang signifikan. Tahap *filtering* juga dilakukan penghapusan *stopwords*. *Stopwords* adalah kata umum (*common words*) yang bisaanya muncul dalam jumlah besar dan dianggap tidak memiliki makna[14]. Contoh *stopwords* adalah "yang", "dan", "di", "dari", dan seterusnya. Contoh dari tahapan ini dapat dilihat pada gambar 2.3.



Gambar 2.3. Proses *Filtering*

2.4. Tokenizing

Tahap *tokenizing* atau parsing adalah tahap pemotongan *string* input berdasarkan tiap kata yang menyusunnya[10]. Pada prinsipnya proses ini adalah memisahkan setiap kata yang menyusun suatu dokumen. Pada umumnya setiap kata terindentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses *tokenizing* mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan kata. Contoh dari tahap ini seperti yang ada dalam Gambar 2.4.



Gambar 4.4. Proses Tokenizing

2.5 Analisis Sentimen

Informasi tekstual secara umum dapat dibagi menjadi informasi fakta dan opini[15]. Fakta adalah ekspresi obyektif terhadap suatu benda, kejadian dan kepunyaan benda tersebut. Opini bisaanya berupa ekspresi subyektif yang menggambarkan sentimen, penilaian, atau perasaan seseorang terhadap suatu benda, kejadian atau kepunyaan dari benda tersebut. [16] menjelaskan bahwa analisis sentimen adalah bagian dari pekerjaan yang meninjau segala sesuatu berhubungan dengan pendapat komputasi, sentimen dan subjektivitas teks. Ditambahkan oleh [17] bahwa analisis sentimen adalah alat untuk memproses koleksi hasil pencarian yang bertujuan dengan mencari atribut suatu produk (kualitas, fitur, dll) dan proses memperolah hasil pendapatnya.

Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, apakah pendapat yang dikemukakan dalam dokumen bersifat positif , negatif atau netral[17]. Penelitian mengenai analisis sentimen telah berkembang sejak tahun 2003 dan merupakan bagian dari *text mining* yang merupakan penelitian komputasi berdasarkan sentimen, *emoticon*, pendapat, komentar dan setiap ekspresi yang diungkapkan oleh teks.

Analisis sentimen difokuskan untuk *review* klasifikasi berdasarkan polaritas. Berdasarkan klasifikasi, analisis sentimen dibagi menjadi dua kelompok utama. Yaitu dokumen klasifikasi ke pendapat atau fakta, atau dikenal sebagai klasifikasi subjektivitas (*subjectivity classification*) dan dokumen klasifikasi ke dalam positif atau negatif, atau dikenal sebaga analisis sentimen. Hal ini adalah proses yang penting untuk menentukan dokumen yang memiliki opini dan dokumen yang menyimpulkan opini bernilai positif, negatif maupun netral.

2.6 Part of Speech Tagging

Part-of-speech Tagging atau yang sering disebut *Tagging* atau *POS Tagging*, merupakan proses pemberian atau penentuan sebuah label terhadap suatu kata dalam suatu kalimat[18]. Sedangkan *part-of-speech* menurut[19] merupakan kategori kata ditinjau dari sudut pandang kebahasaan (gramatikal), seperti kata benda, kata kerja, kata keterangan, kata sifat dan sebagainya.

Ada beberapa pendekatan yang bisa digunakan untuk melakukan *POS Tagging*, yaitu pendekatan berdasar aturan, pendekatan probabilistik, dan pendekatan berbasis transformasi (*transformational based*)[19]. Salah satu perangkat *tagging* yang berbasis metode probabilistik adalah *POS Tagging* untuk bahasa Indonesia yang dibuat oleh Alfan Farizki Wicaksono[19] menggunakan *Hidden Markov Model. Hidden Markov Model (HMM)* adalah sebuah model statistik dari sebuah sistem yang melakukan perhitungan probabilitas dari suatu kejadian yang tidak dapat diamati berdasarkan kejadian yang dapat diamati[9].

Perhitungan probabilitas dilakukan dengan melihat kejadian-kejadian lain yang dapat diamati secara langsung. *HMM POS tagging* memiliki kelebihan dalam memproses *out of vocabulary word*, yaitu kata yang tidak terdapat pada corpus beranotasi. Peran *POS tagging* dalam praproses teks adalah memilah *term-term* yang terbentuk dari suatu kalimat, dalam hal ini suatu *tweet*, berdasarkan kelas kata dalam bahasa Indonesia.. Adapun kelas kata yang dikenal dalam *corpus* bahasa Indonesia ditunjukkan pada tabel 2.1.

Tabel 2.1. Kelas Kata[23]

POS Tag	Arti	Contoh
OP	Kurung Buka	({[
CP	Kurung Tutup	)}]
GM	Garis Miring	/
;	Titik Koma	;
:	Titik Dua	:
“	Tanda Kutip	“
.	Tanda Titik	.
,	Tanda Koma	,
-	Garis	-
…	Tanda Pengganti	…
JJ	Kata Sifat	Baik, Bagus
RB	Kata Keterangan	Sementara, Nanti
NN	Kata Benda	Kursi, Meja
NNP	Benda Bernama	Toyota, Honda
NNG	Benda Berpemilik	Motornya
VBI	Kata Kerja Intransitive	Pergi
VBT	Kata Kerja Transitif	Membeli
IN	Preposisi	Di, Ke, Dari
MD	Modal	Bisa
CC	Kata Sambung Setara	Dan, Atau, Tetapi
SC	Kata Sambung Tidak Setara	Jika, Ketika
DT	Determiner	Para, Ini, Itu
UH	Interjection	Wah, Aduh, Oi
CDO	Kata Bilangan Berurut	Pertama, Kedua, Ketiga
CDC	Kata Bilangan Kolektif	Berdua
CDP	Kata Bilangan Pokok	Satu, Dua, Tiga
CDI	Kata Bilangan Tidak Bisaa	Beberapa
PRP	Kata Ganti Orang	Saya, Mereka
WP	Kata Tanya	Apa. Siapa, Dimana
PRN	Kata Ganti Bilangan	Kedua-Duanya
PRL	Kata Ganti Lokasi	Sini, Situ
NEG	Negasi	Bukan. Tidak
SYM	Symbol	#, %, ^, &, *
RP	Particle	Pun, Kah
FW	Kata Asing	Word

2.8 Rule Based

Metode *Rule Based* ini merupakan metode yang menggunakan aturan bahasa (*grammar*) untuk mendapatkan aturan dalam suatu kalimat[8]. Dalam penelitian ini terdapat proses sebelum analisis sentimen itu sendiri, yaitu *document subjectivity* atau dengan kata lain penentuan suatu *tweet* termasuk opini atau bukan. Menurut [7] untuk menentukan kalimat mana yang

termasuk opini atau bukan, diperlukan *rule* untuk mengolah data hasil proses *POS Tagging*. *Rule* opini yang digunakan dalam penelitian ini ditunjukkan oleh tabel 2.2.

Tabel 2.2 Rule Opini[23]

No	Rule	Contoh
1	RB JJ	Sangat baik, dengan baik, agak baik
2	RB VB	Semoga berjalan, semoga membawa
3	NN JJ	Bukunya bagus, pakaiannya rapi
4	NN VB	Pelajaran membosankan, pekataannya menjengkelkan
5	JJ VB	Mudah difahami, gampang dimaafkan
6	CK JJ	Bagus atau baik, tetapi malas
7	JJ BB	Sama bagus
8	VB VB	Membuat merinding, membikin pusing
9	JJ RB	Indah sekali, bagus sekali
10	VB JJ	Membikin bingung
11	NEG JJ	Tidak seindah, tidak semudah
12	NEG VB	Tidak mengerti, tidak memahami
13	PRP VBI	Saya menyukai
14	PRP VBT	Kita suka
15	VBT NN	Memiliki kedekatan
16	MD VBT	Perlu mengambil referensi
17	MD VBI	Perlu dikembangkan

2.9 Naïve Bayes Classifier

Naïve Bayes Classifier merupakan sebuah metode klasifikasi dengan probabilitas sederhana yang mengaplikasikan Teorema *Bayes* dengan asumsi ketidaktergantungan (independen) yang tinggi. Penggunaan metode *Naive Bayes Classifier* pada penelitian ini didasarkan pada banyaknya dataset yang dipakai sehingga membutuhkan suatu metode yang mempunyai performansi yang cepat dalam pengklasifikasian serta keakuratan yang cukup tinggi[4]. Keuntungan penggunaan *Naive Bayes Classifier* adalah metode ini hanya membutuhkan jumlah data pelatihan (*training data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian[20].

Metode *Naive Bayes Classifier* menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses tehadap sampel data yang sedapat mungkin dapat menjadi representasi data tersebut. Selanjutnya adalah penetuan probabilitas *prior* bagi tiap kategori berdasarkan sampel data. Pada tahap klasifikasi ditentukan nilai kategori dari suatu data berdasarkan *term* yang muncul dalam data yang diklasifikasi. Teorema *Naïve Bayes* dapat dinyatakan dalam persamaan 2.1.

$$P(X_k|\,Y) = \frac{P(Y|X_k)}{\sum_i P(Y|X_i)} \quad \text{.....(2.1)}$$

Dimana, keadaan *Posterior* (Probabilitas X_k di dalam Y) dapat dihitung dari keadaan *prior* (Probabilitas Y di dalam X_k dibagi dengan jumlah dari semua probabilitas Y di dalam semua X_i).

Untuk dapat mengklasifikasikan suatu *tweet*, dalam penelitian ini penulis menggunakan metode *Naive Bayes Classifier* untuk klasifikasi teks, seperti yang dilakukan [21] pada persamaan 2.2.

$$P(v1|C = c) = \frac{\text{CountTerms}(v1\,,\text{docsv}(c))}{\text{AllTerms}(\text{docs}(c))} \quad \text{.....(2.2)}$$

Dimana v1 dalam penelitian ini adalah satu kata tertentu dalam *tweet*, sedangkan $\text{CountTerms}(v1\,,\text{docsv}(c))$ menunjuk pada jumlah kemunculan suatu kata berlabel c ("positif" atau "negatif" atau "netral") . $\text{AllTerms}(\text{docs}(c))$ menunjuk pada jumlah semua kata berlabel c yang ada pada dataset. Untuk menghindari adanya nilai nol pada probabilitas, maka diberlakukan *Laplace (add-one) smoothing*. Tujuan daripada *smoothing* adalah untuk mengurangi probabilitas dari hasil/keluaran yang terobeservasi, dan juga sekaligus meningkatkan/menambah probabilitas hasil/keluaran yang belum terobservasi[22], sehingga persamaan menjadi sebagi berikut:

$$P(v1|C = c) = \frac{\text{CountTerms}(v1\,,\text{docsv}(c))+1}{\text{AllTerms}(\text{docs}(c))+|V|} \quad \text{.....(2.3)}$$

Dimana $|V|$ menunjuk pada jumlah semua kata dalam tweet yang ada di dataset.

3. METODOLOGI PENELITIAN

3.1. Studi Literatur dan Pemahaman

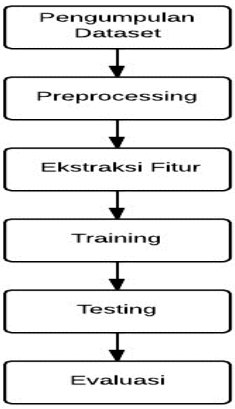
Pada tahap ini, penulis mencari dan mempelajari referensi berupa textbook, artikel ilmiah maupun jurnal yang berkaitan dengan penelitian. Topik yang akan dibahas antara lain : analisis sentimen, *part-of-speech tagging* pada kalimat bahasa Indonesia dan metode klasifikasi *Naive Bayes Classifier*.

3.2. Pengumpulan Data

Dataset yang akan digunakan dalam penelitian adalah public timeline tweet bahasa Indonesia yang merupakan hasil pencarian berdasarkan percakapan seseorang terhadap account resmi tokoh politik yang akan maju sebagai calon presiden Indonesia 2014. Beberapa account resmi yang diambil diantaranya, *@jokowi_do2* (Joko Widodo), *@Prabowo08* (Prabowo Subianto), *@Pak_JK* (Jusuf Kalla), *@hattarajasa* (Hatta Rajasa). Dataset didapatkan dengan cara *streaming* memanfaatkan API dari Twitter kemudian disimpan dalam bentuk database. Pengambilan data tweet dilakukan pada tiga periode berbeda. Yang pertama periode sebelum pemilu legislatif, yaitu pada tanggal 1 Februari sampai 3 Februari 2014. Kemudian saat pemilu legislatif, antara tanggal 7 April sampai 9 April 2014. Periode terakhir adalah setelah pengumuman deklarasi pasangan capres dan cawapres yang resmi maju ke pemilihan umum presiden dan wakil presiden Indonesia tahun 2014.

3.3. Implementasi

Implementasi penelitian ini dilakukan dengan langkah kerja seperti gambar. 3.1.

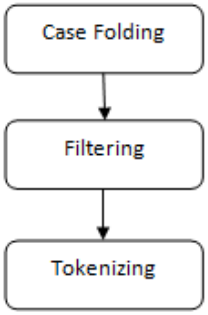


Gambar 3.1 Diagram alir kerja penelitian

Proses awal pada tahapan implementasi adalah pengumpulan dataset yang akan digunakan baik untuk *testing* maupun *training*. Pengumpulan dasaset dengan cara memanfaatkan API dari pihak Twitter melalui teknik *streaming* data. Hasil dari *streaming* data kemudian disimpan kedalam database MySQL. Kemudian dilakukan *preprocessing data* untuk menjadikan sederhana dimensi dari dataset. Selanjutnya dataset tersebut kembali dipilah melalu ekstraksi fitur dengan cara memanfaatkan *part-of-speech tagging* untuk mendapatkan kelas kata. Setelah mendapatkan kelas kata dari dataset, dilakukan filtering opini menggunakan *rule based* yang dibangun oleh penelitian sebelumnya[7] untuk menentukan suatu tweet termasuk opini atau bukan. Hasil dari tweet yang sudah masuk kategori opini akan dipisahkan menjadi 3 kategori sentimen. Yaitu sentimen yang mempunyai polaritas positif, negatif dan netral. Hasil *tweet* yang dipisahkan tadi nantinya akan digunakan sebagai dataset *training* dimana proses pemisahannya dilakukan secara manual. Setelah dataset training terbentuk, barulah klasifikasi *tweet* menggunakan metode *Naive Bayes Classifier* dimulai.

3.3.1 Preprocessing Data

Sebelum dataset tweet siap digunakan maka akan terlebih dahulu dilakukan *preprocessing data* sehingga dataset telah bersih dan siap digunakan dalam proses selanjutnya. Berikut adalah diagram alir mengenai *preprocessing data*.



Terlihat dari diagram diatas bahwa tahap pertama *preprocessing data* adalah *case folding*, yaitu mengubah semua huruf kapital menjadi huruf kecil atau *lowercase*. Kemudian dilakukan *filtering* berupa penghapusan semua karakter selain string serta penghapusan beberapa karakteristik dari data twitter, misalnya *@username*, *#hashtag*, http://*URL*, dan "RT" atau kata yang menandakan kalau itu perulangan tweet. Dalam filtering ini juga dilakukan penghapusan terhadap *stopword*. Hal ini bermanfaat untuk mengurangi *load* atau *performance* saat melakukan training maupun testing dataset. Selanjutnya dataset akan dilakukan *tokenizing*, yaitu pemecahan berdasarkan perkata. Hal ini bisa dilakukan dengan menandai karakter spasi sebagai pembatas.

3.2.2. Ekstraksi Opini

Pada tahap ini akan dilakukan penyaringan terhadap tweet yang telah dilakukan *preprocessing data* sebelumnya untuk mendapatkan tweet yang hanya mengandung opini / sentimen. Tahap penyaringan dimulai dari proses *part-of-speech tagging* menggunakan alat *HMM POS Tagging* terhadap dataset *tweet*, selanjutnya dataset difilter menggunakan *rule based* yang menunjukkan kalimat opini dengen memakai *rule* opini dibangun pada penelitian sebelumnya. *Rule based* opini dapat dilihat pada table 2.2.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Penelitian

Hasil dari peneltian ini dapat dibagi menjadi enam bagian. Yaitu hasil pengujian akurasi metode *Naïve Bayes Classifier*, hasil persebaran tweet, hasil analisis sentimen terhadap calon presiden, hasil analisis sentimen terhadap calon wakil presiden hasil analisis sentimen terhadap pasangan calon presiden dan wakil presiden, dan terakhir adalah perbandingan dengan data aktual.

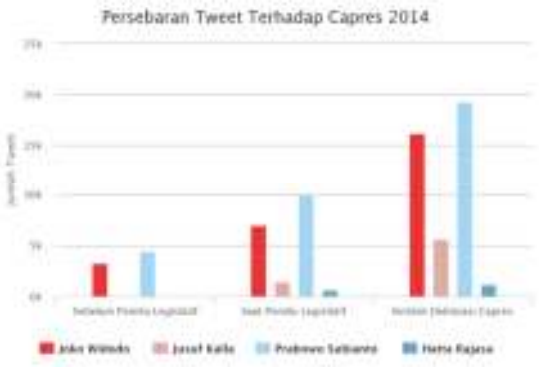
4.1.1. Pengujian Naïve Bayes Classifier

Pengujian akurasi metode naïve bayes classifier dilakukan dengan cara eksperimen Tingkat akurasi dari metode naïve bayes classifier untuk klasifikasi polaritas sentimen tweet terhadap calon presiden Indonesia 2014 dapat dihitung dengan persamaan 4.1

$$\text{Akurasi} = \frac{\text{jumlah prediksi benar}}{\text{jumlah data}} \text{x} 100\% \quad \text{.....(4.1)}$$

Hasil yang didapatkan dari pengujian 100 data random yang sudah diklasifikasikan polaritas secara manual dengan menggunakan 1400 data training mendapatkan akurasi sebeasar 90%.

4.1.2. Hasil Persebaran Tweet



4.1 Statistik Persebaran Tweet Capres-Cawapres Indonesia Tahun 2014

Grafik 4.1 menunjukkan persebaran opini masyarakat melalui jejaring sosial Twitter terhadap *account* resmi masing-masing tokoh politik yang akan maju sebagai calon presiden dan calon wakil presiden Indonesia tahun 2014. Berdasarkan dari grafik di atas, terlihat pada periode waktu sebelum pemilu legislatif jumlah percakapan yang membicarakan *account* resmi dari

Prabowo Subianto lebih banyak dari tokoh lainnya. Prabowo Subianto unggul sebanyak 4474 *tweet*, dibandingkan dengan Joko Widodo yang mendapat 3281 *tweet*. Sedangkan Jusuf Kalla dan Hatta Rajasa hanya mendapat 35 dan 8 *tweet* saja. Kemudian pada periode waktu selanjutnya, yaitu saat diadakan pemilu legislatif, perolehan *tweet* Prabowo Subianto masih unggul dibandingkan tokoh politik yang lainnya dengan perolehan 9942. Periode yang terakhir adalah saat setelah deklarasi pengumuman pasangan capres dan cawapres yang resmi akan maju dalam pemilu Indonesia tahun 2014 menunjukkan bahwa Prabowo Subianto tetap unggul dibanding tokoh politik yang lain. Prabowo Subianto unggul mendapat 19297 *tweet*, Joko Widodo mendapat 16212 *tweet*, Jusuf Kalla mendapat 5647 tweet dan Hatta Rajasa mendapat 1213 *tweet*.

Hal ini dapat disimpulkan bahwa percakapan atau opini masyarakat terhadap account resmi tokoh politik yang akan maju sebagai calon presiden dan wakil presiden Indonesia tahun 2014, menempatkan Prabowo Subianto sebagai capres yang banyak dibicarakan di jejaring sosial Twitter. Sedangkan untuk cawapres, Jusuf Kalla yang menjadi pasangan dari Joko Widodo mendapat opini lebih tinggi dari Hatta Rajasa dimana beliau adalah cawapres dari Prabowo Subianto. Dari grafik ini pula dapat dilihat bahwa perkembangan antara periode pengambilan data *tweet* semakin naik pada setiap periode waktu.

4.1.3. Analisis Sentimen Terhadap Calon Presiden

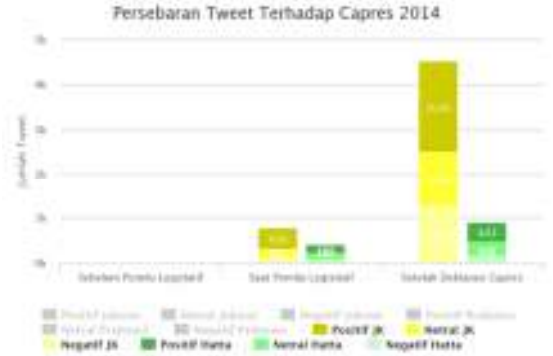
4.2 Statistik Polaritas Tweet Capres Indonesia Tahun 2014

Grafik 4.2 terlihat bahwa kecenderungan polaritas sentimen positif berfluktuasi. Periode pertama, yaitu saat sebelum pemilu legislatif, Prabowo Subianto mendapatkan sentimen positif sebesar 43% dan 18,6% sentimen negatif. Kemudian saat pemilu legislatif terjadi peningkatan sentimen positif dan sentimen negatif menjadi 54% dan 31%. Terakhir saat setelah deklarasi dengan calon wakil presiden Hatta Rajasa terjadi penurunan sentimen positif menjadi 45,2%. Hal ini juga dibarengi dengan penurunan sentimen negatif menjadi 25,6%.

Kecenderungan polaritas sentimen positif terhadap Joko Widodo juga berfluktuasi. Periode pertama saat sebelum pemuli legislatif, Joko Widodo hanya mendapat 24,6%. Setelah pemilu legislatif terjadi peningkatan sentimen positif terhadap Joko Widodo menjadi 37%, akan tetapi hal ini juga dibarengi dengan peningkatan sentimen negatif terhadap Joko Widodo menjadi 41,2%. Kemudian saat setelah deklarasi dengan pasangan calon wakil presiden Jusuf Kalla terjadi penurunan sentimen positif menjadi 36,5%. Akan tetapi pada periode ini, terjadi penurunan sentimen negatif terhadap Joko Widodo mejadi 34,4%.

Perolehan sentimen positif pada keseluruhan periode waktu menunjukkan calon presiden Prabowo Subianto lebih unggul mendapatkan 47,6% dibandingkan calon presiden Joko Widodo yang mendapatkan 35%. Sedangkan untuk sentimen negatif, Joko Widodo mendapatkan perolehan yang lebih tinggi sebesar 36,9% dibandingkan dengan Prabowo Subianto yang mendapat sentimen negatif sebesar 26,9%. Dan untuk sentimen berpolaritas netral Joko Widodo lebih unggul sebanyak 36,9% sedangan Prabowo Subianto mendapatkan 26,9%.

4.1.4. Analisis Sentimen Terhadap Calon Wakil Presiden



4.3 Statistik Polaritas Tweet Cawapres Indonesia Tahun 2014

Grafik 4.3 adalah hasil polaritas sentimen terhadap tokoh politik yang menjadi calon wakil presiden Indonesia pada pemilihan umum tahun 2014. Hatta Rajasa sebagai calon wakil presiden dari Prabowo subianto pada keseluruhan periode waktu mendapatkan sentimen positif yang lebih tinggi daripada Jusuf Kalla yang menjadi calon wakil presiden dari Joko Widodo Hatta Rajasa mendapat 48,6% sedangkan Jusuf Kalla menyusul dengan selih tipis sebesar 47,2%. Dilihat dari sentimen negatif, Jusuf Kalla mendapatkan sentimen negatif yang lebih tinggi daripada Hatta Rajasa. Jusuf Kalla mendapatkan 25,4% sedangkan Hatta Rajasa mendapatkan 16,3%. Sedangkan untuk sentimen berpolaritas netral, perolehan Jusuf Kalla lebih tinggi dibandingkan dengan Hatta Rajasa, yaitu 25,4% dan 16,3%. Untuk rincian persebaran polaritas sentimen dapat dilihat pada grafik 4.3.

Grafik 4.3 menggambarkan bagaimana polaritas sentimen terhadap tokoh politik yang akan maju sebagai calon wakil presiden Indonesia 2014. Kecenderungan sentimen terhadap calon wakil presiden berfluktuasi. Pada periode pertama, yaitu sebelum pemilu legislatif, terlihat masih sedikit jumlah sentimen. Jusuf Kalla yang menjadi calon wakil presiden dari Joko Widodo pada eriode pertama menunjukkan total percakapan sebanyak 24 tweet dengan nilai polaritas masing-masing 14 untuk sentimen positif, 6 sentimen negatif dan 4 sentimen netral. Sedangkan Hatta Rajasa sebagai calon wakil presiden dari Prabowo subianto mendapatkan total percakapan 6 dengan rincian 3 sentimen negative dan 3 sentimen netral, tidak memiliki sentimen berpolaritas positif. Kemudian pada periode kedua, yaitu saat pemilu legislatif 2014, terjadi peningkatan sentimen. Kecenderungan sentimen positif terhadap Jusuf Kalla

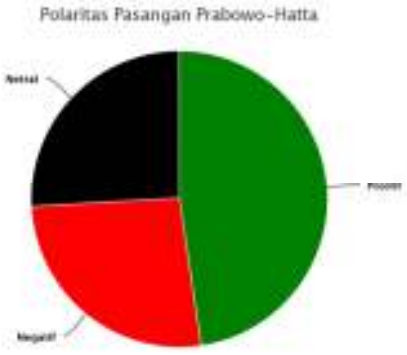
adalah 60,4%. Sedangkan kecenderungan sentimen positif tehadap Hatta Rajasa pada periode kedua adalah 48%. Kemudian periode ketiga yaitu saat setelah deklarasi berpasangan dengan calon presiden, entimen positif terhadap Jusuf Kalla cenderung menurun menjadi 44,8%. Sedangkan Hatta Rajasa terjadi peningkatan sentimen positif menjadi 49, 2%.

Secara keseluruhan periode waktu, Hatta Rajasa mendapatkan sentimen positif yang lebih tinggi daripada Jusuf Kalla. Hatta Rajasa mendapat 48,6% sedangkan Jusuf Kalla menyusul dengan selisihh tipis sebesar 47,2%. Dilihat dari sentimen negatif, Jusuf Kalla mendapatkan sentimen negatif yang lebih tinggi daripada Hatta Rajasa. Jusuf Kalla mendapatkan 25,4% sedangkan Hatta Rajasa mendapatkan 16,3%. Sedangkan untuk sentimen berpolaritas netral, perolehan Jusuf Kalla lebih tinggi dibandingkan dengan Hatta Rajasa, yaitu 25,4% dan 16,3%.

4.1.5. Analisis Sentimen Terhadap Pasangan Calon Presiden dan Wakil Presiden

4.4 Statistik Polaritas Pasangan Joko Widodo – Jusuf Kalla

Grafik 4.4 mununjukkan hasil polaritas sentimen terhadap capres dan cawapres Joko Widodo dan Jusuf Kalla. Terlihat sentimen positif terhadap pasangan Joko Widodo – Jusuf Kalla sebesar 37,6%. Hal ini menunjukkan peningkatan terhadap sentimen positif Joko Widodo yang sebelumnya mendapatkan 35% sebelum berpasangan dengan Jusuf Kalla. Sedangkan polaritas sentimen negatif terjadi penurunan setelah Joko Widodo berpasangan dengan Jusuf Kalla. Penurunan terjadi dari semula 36,9% menjadi 34,4%. Penurunan juga terjadi pada polaritas sentimen netral pasangan Joko Widodo – Jusuf Kalla yaitu sebesar 0,2% dari 28,1% menjadi 27,9%.



4.5 Statistik Polaritas Pasangan Prabowo Subianto-Hatta Rajasa

Grafik 4.5 mununjukkan hasil polaritas sentimen terhadap capres dan cawapres Prabowo Subianto - Hatta Rajasa. Sentimen terhadap Prabowo Subianto setelah berpasangan dengan Hatta Rajasa terjadi kenaikan tipis dari 47,6% menjadi 47,7%. Sedangkan sentimen berpolaritas negatif juga terjadi penurunan sebesar 0,5% dari 26,9% menjadi 26,4%. Untuk sentimen berpolaritas netral, terjadi kenaikan dari 25,2% menjadi 25,9%.

5. KESIMPULAN

Hasil dari pengamatan sentimen masyarakat melalui jejaring sosial Twitter menunjukkan jumlah percakapan terhadap capres dan cawapres di jejaring sosial Twitter menjelang mendekati pemilu 2014 semakin meningkat. Jumlah percakapan terhadap pasangan Prabowo Subianto – Hatta Rajasa lebih unggul sebesar 53% dibandingkan percakapan terhadap pasangan Joko Widodo – Hatta Rajasa yang mendapat 47 %.

Hasil dari pengamatan polaritas sentiment masyarakat terhadap pasangan calon presiden dan wakil preside menunjukkan pasangan Prabowo Subianto – Hatta Rajasa mendapatkan 47,7% untuk sentimen positif, 26,4% sentimen negatif dan 25,9% sentimen netral. Sedangkan pasangan Joko Widodo – Jusuf Kalla mendapatkan total 37,6% sentimen positif, 34,4% sentimen negatif, dan 27,9 sentimen netral. Dari hal ini dapat disimpulkan bahwa pasangan Prabowo Subianto – Hatta Rajasa lebih unggul dari pasangan Joko Widodo – Jusuf Kalla dalam hal jumlah percakapan dan sentimen positif pada jejaring sosial Twitter.

6. SARAN

Adapun saran yang dipertimbangkan untuk pengembangan penelitian selanjutnya adalah dengan mengkombinasikan sistem deteksi spam *tweet* untuk meningkatkan kualitas hasil *tweet*. Kemudian pada penelitian selanjutnya bisa menggunakan data opini dari jejaring sosial selain Twitter seperti Facebook, Youtube, Path, komentar bertita, maupun forum untuk mengetahui perbandingan bagaimana keakuratan hasil analisa sentimen dengan data aktual yang tejadi.

7. DAFTAR PUSTAKA

[1] Beritagar.com. (2013). Statistik pengguna Twitter Indonesia Oktober 2013. Diakses 30 Desember 2013. (http://beritagar.com/p/statistik-pengguna-twitter-indonesia-oktober-2013-10207)

[2] Bolen, Johan; Mao, Huina; Zeng, Xiaojung. (2011) *Twitter mood predicts the stock market. Dalam Journal of Computational Science 2* p.1–8

[3] Natalius, Samuel., 2011, Metoda *Naïve Bayes Classifier* dan Penggunaannya pada Klasifikasi Dokumen.

[4] Larose, D. T. 2006. *Naïve Bayes Estimation and Bayesian Networks, in Data Mining Methods and Models*, John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/0471756482.ch5

[5] McCue, Rita. (2009) *A Comparison of the Accuracy of Support Vector Machine and Naıve Bayes Algorithms In Spam Classification*.

[6] Parikh, R., & Movassate, M. (2009). Sentimen Analysis of User Generated Twitter Updates using Various Classification

[7] Rozi, Imam F; Pramono Sholeh H; Dahlan, Achmad E. (2012) Implementasi Opinion

Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi. Dalam Jurnal EECCIS Vol. 6, No. 1, Juni 2012.

[8] Hearst, M. A. (1997) *Text data mining: Issues, techniques, and the relationship to information access*. Presentation notes for UW/MS workshop on data mining, July 1997

[9] Feldman, R & Dagan, I. (1995*) Knowledge discovery in textual databases (KDT). Dalam Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (KDD-95), Montreal, Canada, August 20-21, AAAI Press, 112-117

[10] Feldman, R & Sanger, J. (2007) *The Text Mining Handbook-Advanced Approaches in Analyzing Unstructured Data*, USA: New York.

[11] Triawati, Candra; Bijaksana, M.Arif; Indrawati, Nur; Saputro, Widyanto Adi. (2009) Pemodelan Berbasis Konsep Untuk Kategorisasi Artikel Berita Berbahasa Indonesia. Dalam Seminar Nasional Aplikasi Teknologi Informasi 2009.

[12] Go, Alec; Bhayani, Richa; Huang, Lei. (2009) *Twitter Sentimen Classification using Distant Supervision.*

[13] Pak, A., dan Paurobek, P., 2010, *Twitter as a Corpus for Sentimen Analysis and Opinion Mining*, Universite de Paris-Sud, Laboratoire LIMSI-CNRS. Batiment 508, F-91405 Orsay Cedex, France.

[14] Yates, Baeza R. & Neto, Ribero B. (1999) *Modern Information Retrieval. New York*: ACM Press.

[15] B. Liu. (2010) *Handbook of Natural Language Processing, chapter Sentimen Analysis, 2nd Edition.*

[16] Pang, Bo & Lilian, Lee., 2008, *Opinion Mining and Sentimen Analysis*. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135

[17] Dave, Kushal; Lawrence, Steve; Pennock, David M. (2003) *Mining the peanut gallery: Opinion Extraction and semancitc classification of product reviews*.

[18] Jurafsky, Daniel & Martin, H. James. (2007) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Lingustics, and Speech Recognition*.

[19] Rozi, Imam F. Implementasi *Rule-Based Document Subjectivity* Pada Sistem *Opinion Mining*. Dalam Jurnal ELTEK, Vol 11 No 01, April 2013.

[20] Kao, A & Poteet, S. (2007) *Natural Language Processing and Text Mining*. Springer Verlag, London, England

[21] Ricci, F.; Rokach, L.; et al. 2011. Recommender Systems Handbook. Berlin : Springer.

[22] Arguello, J., 2013. *Naïve Bayes Text Classification* (https://ils.unc.edu/courses/ 2013_fall/inls613_001/lectures/04NaiveBayes Classification.pdf. Diakses tanggal 12 April 2014). The University of North Carolina.

[23] Beritasatu.com. (2014). Jokowi-JK, Presiden dan Wapres Pilihan Netizen. Diakses 15 Agustus 2014. (http://www.beritasatu.com/nasional/195024-jokowijk-presiden-dan-wapres-pilihan-netizen.html)