

## Analisis Sentimen Data Twitter Tentang Pasangan Capres-Cawapres Pemilu 2019 Berbasis Metode Lexicon Dan Support Vector Machine

Danar Wido Seno<sup>1</sup>, Arief Wibowo<sup>2</sup>

<sup>1,2</sup>Magister Ilmu Komputer, Universitas Budi Luhur

Jl. Ciledug Raya, RT.10/RW.2, Petukangan Utara, Kebayoran Lama Jakarta Selatan 12260

Telp. (021) 5853753 ext. 227, Fax. (021) 5869225

E-Mail : danarwidoseno@gmail.com<sup>1</sup>, arief.wibowo@budiluhur.ac.id<sup>2</sup>

### Abstract

*Social media writing content growing make a lot of new words that appear on Twitter in the form of words and abbreviations that appear so that sentiment analysis is increasingly difficult to get high accuracy of textual data on Twitter social media. In this study, the authors conducted research on sentiment analysis of the pairs of candidates for President and Vice President of Indonesia in the 2019 Elections. To obtain higher accuracy results and accommodate the problem of textual data development on Twitter, the authors conducted a combination of methods to conduct the sentiment analysis with unsupervised and supervised methods. namely Lexicon Based. This study used Twitter data in October 2018 using the search keywords with the names of each pair of candidates for President and Vice President of the 2019 Elections totaling 800 datasets. From the study with 800 datasets the best accuracy was obtained with a value of 92.5% with 80% training data composition and 20% testing data with a Precision value in each class between 85.7% - 97.2% and Recall value for each class among 78, 2% - 93.5%. With the Lexicon Based method as a labeling dataset, the process of labeling the Support Vector Machine dataset is no longer done manually but is processed by the Lexicon Based method and the dictionary on the lexicon can be added along with the development of data content on Twitter social media.*

*Keywords: Sentiment Analysis, Support Vector Machine, Lexicon Based, Twitter Sentiment Analysis, Sentiment Analysis with R Programming.*

### Abstrak

*Semakin berkembangnya tulisan pada media sosial membuat banyak kata-kata baru yang bermunculan dalam twitter baik berupa kata-kata maupun singkatan yang bermunculan sehingga analisis sentimen semakin sulit untuk mendapatkan akurasi yang tinggi terhadap data textual pada media sosial twitter. Pada penelitian ini penulis melakukan riset mengenai analisis sentiment terhadap pasangan calon Presiden dan Wakil Presiden Indonesia pada Pemilu 2019. Untuk mendapatkan hasil akurasi yang lebih tinggi dan mengakomodir masalah perkembangan data textual pada twitter penulis melakukan kombinasi metode untuk melakukan analisis sentiment tersebut dengan metode unsupervised dan supervised yaitu Lexicon Based. Studi ini menggunakan data twitter pada bulan Oktober 2018 dengan menggunakan keyword pencarian berdasarkan nama masing-masing pasangan calon Presiden dan Wakil Presiden Pemilu 2019 sebanyak 800 dataset. Dengan 800 dataset tersebut akurasi terbaik didapat dengan nilai 92,5% dengan komposisi data training 80% dan data testing 20% dengan nilai Precision pada setiap kelas diantara 85,7% - 97,2% dan nilai Recall setiap kelas diantara 78,2% - 93,5%. Dengan metode Lexicon Based sebagai pelabelan dataset dapat membuat proses pelabelan dataset Support Vector Machine tidak lagi dilakukan secara manual melainkan diproses oleh metode Lexicon Based dan dictionary pada lexicon dapat ditambahkan seiring dengan perkembangan konten data pada media sosial twitter.*

*Kata kunci: Sentiment Analysis, Support Vector Machine, Lexicon Based, Twitter Sentiment Analysis, Sentiment Analysis dengan R Programming.*

## I. PENDAHULUAN

Twitter adalah salah satu diantara media sosial yang hingga saat ini masih digunakan oleh orang-orang untuk menuliskan opini maupun emosional yang mereka rasakan dan menuliskannya di media twitter. Tweet atau kicauan pada twitter bisa berupa pendapat, saran, ataupun kritikan tentang topik-topik tertentu. Semakin berkembangnya waktu isi kicauan dalam twitter kini semakin banyak mendatangkan kata-kata baru, istilah-istilah baru yang muncul dalam kicauan/*tweet* serta penggunaan bahasa yang tidak baku pada tweet sehingga untuk mendapatkan akurasi yang tinggi pada analisis sentimen menjadi lebih sulit dilakukan. Pada studi ini data tweet yang dijadikan studi kasus adalah data opini masyarakat terhadap pasangan calon Presiden dan Wakil Presiden Indonesia pada Pemilu tahun 2019, yaitu Jokowi-Ma'ruf dan Prabowo-Sandi. Dari Pemaparan latar belakang masalah yang dihadapi terdapat peluang untuk melakukan optimasi akurasi dengan kombinasi metode pada sentiment analysis. Kombinasi metode yang akan dilakukan pada penelitian ini adalah menggunakan pendekatan berbasis *knowledge-based* dan *machine learning-based*, dimana dari penelitian sebelumnya yang dilakukan oleh Nomleni pada tahun 2015 dengan menggunakan metode *Support Vector Machine* mendapatkan akurasi terbaik sebesar 80%, dan penelitian dengan metode *Holistic Lexicon-Based* yang dilakukan oleh Purba, Hidayati dan Gozali pada tahun 2014 mendapatkan akurasi sebesar 88,6%. Kelemahan dari pendekatan *knowledge-based* adalah terbatasnya *dictionary*, menurut Nurfalih (dalam Matulatuwa, Sediyo dan Iriani, 2017) mengatakan bahwa "kamus/dictionary adalah komponen penting dalam sistem yang menggunakan metode *Lexicon Based*". Pendekatan *machine learning* dapat memberikan kontribusi terhadap metode *knowledge-based* dapat membantu pendekatan *machine-learning* untuk melakukan pelabelan dataset latih pada metode *Support Vector Machine* seperti yang dilakukan oleh Imam Syaefi dan Hendri Murfi pada tahun 2014, namun identifikasi tweet nya menggunakan SentiWordNet dan hasil tweet diterjemahkan kedalam bahasa inggris. Atas dasar teori-teori tersebut penulis mengusulkan sebuah kombinasi metode untuk analisis sentimen menggunakan *Lexicon-Based* dan *Support Vector Machine* dengan harapan kombinasi metode dan algoritma tersebut dapat saling berkontribusi terhadap proses pelabelan sentimen untuk mendapatkan akurasi yang lebih baik.

## II. TINJAUAN PUSTAKA

### a. Media Sosial

Media Sosial adalah merupakan media yang bermanfaat sebagai sarana bersosialisasi antar pengguna secara online tanpa dibatasi ruang dan waktu. Pengertian media sosial menurut David Meerman Scott [1] bahwa media sosial menyediakan tempat dan cara untuk orang berbagi ide, konten, pemikiran, dan hubungan ketika kita sedang terhubung dengan internet atau dunia maya. Dengan demikian maka batasan ruang dan waktu menjadi tereliminasi di antara pengguna-pengguna media sosial.

### b. Twitter

Twitter adalah layanan jejaring sosial yang populer digunakan untuk menyampaikan dan membaca pesan berbasis teks sepanjang 140 karakter, yang akhirnya mampu melayani hingga 280 karakter. Istilah pesan Twitter dikenal dengan sebutan kicauan. Menurut Bolen (dalam Nurhuda dan Sihwi, 2014) [6] menyatakan bahwa "Tweet adalah teks status pengguna yang digunakan untuk memberikan informasi melalui Twitter".

### c. Text Mining

Menurut Nurhuda dan Sihwi, *Text mining* didefinisikan sebagai suatu proses menemukan informasi dari sekumpulan dokumen berupa data tekstual menggunakan metode analisis tertentu, proses ini masih merupakan bagian dari data mining" [6]. *Text mining* bertujuan untuk menemukan informasi yang lebih bermanfaat dari sekumpulan dokumen, atau sekumpulan teks dengan format yang tidak terstruktur atau semi terstruktur. Beberapa pekerjaan text mining yang dapat diselesaikan adalah pengkategorisasian atau pengelompokan teks. Tahapan-tahapan yang umumnya dilakukan pada text mining antara lain: *Tokenizing*, *Filtering*, *Stemming*, *Tagging* dan *Analyzing*.

### d. Analisis Sentimen

Analisis sentimen adalah sebuah riset secara komputasi pada opini atau emosi yang diekspresikan secara tekstual oleh seseorang atau kelompok [2]. Analisis sentimen merupakan turunan proses dari *text mining*, yang secara khusus melakukan analisis dokumen teks atau data tekstual. Umumnya proses analisis dokumen teks akan mengelompokkan sentimen dalam wujud positif, negatif maupun netral.

### e. Support Vector Machine

*Support Vector Machine* (SVM) merupakan salah satu pengklasifikasi diskriminatif dengan *hyperplane* pemisah. Metode ini melakukan pelatihan data berlabel (*supervised learning*), sementara itu, algoritma menghasilkan *hyperplane* optimal yang mengkategorikan contoh baru. *Hyperplane* adalah garis yang memisahkan sebuah *plane* menjadi dua bagian pada setiap kelas yang terletak di kedua sisi. Menurut Santoso, 2007 [3] *Support vector machine* (SVM) merupakan salah satu teknik terbaik untuk menyelesaikan proses

prediksi, dalam kasus klasifikasi atau regresi. Berikut adalah rumus perhitungan hyperplane pada Support Vector Machine:

$$f(x) = \mathcal{W} \cdot \mathcal{X} + b \quad (2.1)$$

Sumber: [4]

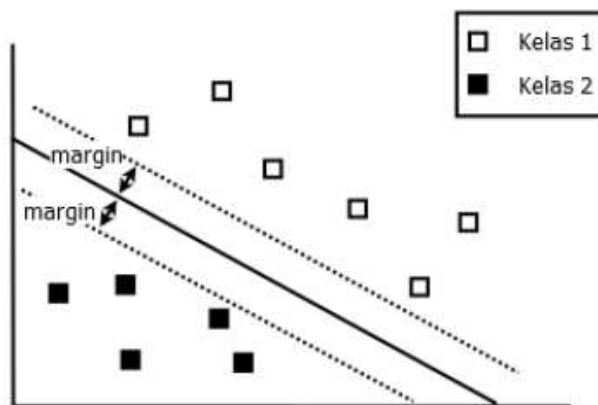
Dimana :

W = parameter *hyperplane* yang dicari (garis yang tegak lurus antara garis *hyperplane* dan titik support vector)

x = titik data masukan *Support Vector Machine*

b = parameter *hyperplane* yang dicari (nilai bias)

f = fungsi *hyperplane*



Gambar 1. *Hyperplane* Memisahkan 2 Kelas [5]

#### 1. Linear Kernel SVM

Linear kernel adalah fungsi kernel paling sederhana, dengan bentuk persamaan linear sebagai berikut:

$$K(x, y) = x^T y + c \quad (1)$$

Sumber: [7]

#### 2. Polynomial Kernel

Polinomial kernel adalah fungsi kernel untuk data yang tidak terpisah secara linear, dengan bentuk persamaan sebagai berikut:

$$K(x, y) = (x^T y + c)^d \quad (2)$$

Sumber: [7]

#### 3. Radial Basis Function (RBF) Kernel

RBF kernel adalah fungsi kernel untuk analisis pada data yang diketahui tidak terpisah secara linear, dengan dua jenis parameter yaitu *Gamma* dan *Cost* dengan bentuk persamaan sebagai berikut:

$$K(x, y) = \exp\left(\frac{-||x-y||^2}{2\sigma^2}\right) \quad (3)$$

Sumber: [7]

#### f. *Lexicon Based*

Pendekatan *Lexicon-Based* menggunakan *dictionary* atau kamus *lexicon* untuk melakukan penilaian terhadap kata. Pada *dictionary*, kata-kata dipasangkan dengan nilai polaritasnya. Yang harus dilakukan sebelum melakukan analisis menggunakan *lexicon* adalah menentukan kata yang akan dianalisis dari korpus.. Menurut Melville dan Lawrence [7] menyatakan bahwa *Lexicon Based* didasarkan pada asumsi bahwa orientasi sentimen kontekstual merupakan jumlah dari orientasi sentimen setiap frase atau kata. Berikut adalah contoh kamus dan isinya:

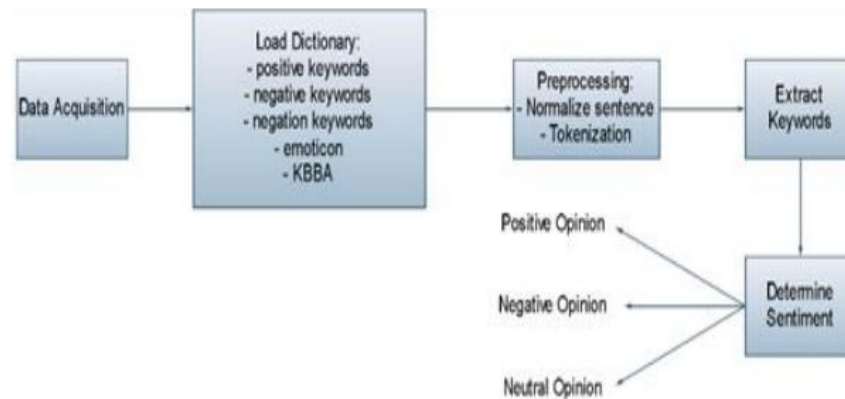
1. *Positive keywords*: banyak, baik bisa, ok, *best*, pintar, lancar, cepat, bagus, senang.

2. *Negative keywords*: bangkrut, banjir, bodoh, gagal, kurang, susah, lambat, parah, bohong.

3. *Negation keywords*: terdiri dari frasa belum, tidak, bukan

4. *Emoticon*: misalnya *smile* :-) yang diberikan nilai 1, atau *sad* :-( yang dapat diberikan nilai -1

5. Kamus konversi bahasa gaul seperti '*bgm*' untuk bagaimana, '*bgs*' untuk bagus, atau '*beud*' untuk banget



Gambar 2. Algoritma Metode *Lexicon Based* [7]

1. *Data Acquisition*, merupakan tahap penarikan data dari media sosial, dengan luaran berupa daftar opini pembaca serta metadata seperti: *username*, lokasi, ID, waktu.
2. *Load Dictionary*, merupakan proses pemanggilan kamus, berisi kata kunci yang menunjukkan sentimen positif, negatif, netral, atau penemuan bahasa gaul pada frasa.
3. *Preprocessing*, merupakan tahap penyiapan kalimat sebelum proses ekstraksi kata kunci dan penentuan sentimen, umumnya terdiri dari normalisasi kalimat maupun proses tokenisasi.
4. *Extract Keyword*, merupakan tahap untuk mengekstraksi kata kunci yang akan menjadi penentu sentimen positif dan negatif atau netral.
5. *Determine Sentiment*, merupakan tahap menentukan sentimen dari suatu kalimat opini, yang dilakukan dengan melakukan perhitungan probabilitas kemunculan kata kunci positif dan kata kunci negatif.

#### g. K-fold Cross Validation

*Cross validation* merupakan salah satu metode statistik untuk mengevaluasi kinerja model atau algoritma, metode ini bekerja dengan cara memisahkan dua subset data, yang terdiri dari pembelajaran dan evaluasi. Adapun manfaat dari metode ini, adalah dilakukannya proses berulang secara acak *sub-sampling* sehingga semua pengamatan digunakan untuk proses pelatihan dan validasi, selain itu setiap pengamatan dipastikan telah digunakan untuk proses validasi [4].

#### h. Tinjauan Studi

Tinjauan studi dilakukan dengan mengkaji paper dan jurnal ilmiah dari penelitian-penelitian sebelumnya yang terkait dengan penelitian ini. Pendekatan dengan machine learning memiliki tingkat akurasi rata-rata diangka 80% sampai dengan 90% seperti metode *Naive Bayes* dan *Support Vector Machine*. Dalam penelitian yang menggunakan pendekatan knowledge-based seperti *Lexicon-Based* yang dilakukan oleh Kundi FM, dkk dapat mendapatkan presentase 92% dalam klasifikasi biner namun perlu meningkatkan *precision* dalam kasus negatif dan dalam kasus netral. Dari beberapa penelitian diatas, metode *Lexicon-Based* menjadi salah satu metode yang banyak dilakukan modifikasi dan kombinasi dan memiliki peluang untuk melakukan peningkatan hasil akurasi dalam sentiment analysis. Pada penelitian ini akan digunakan kombinasi metode *Lexicon-Based* dan *Support Vector Machine*, masing-masing dari metode tersebut akan memberikan kontribusi terhadap metode lainnya untuk peningkatan *accuracy*, *precision* dan *recall* terhadap hasil klasifikasi *sentiment analysis* pada penelitian ini.

### III. METODE PENELITIAN

Studi ini merupakan eksperimen dari kombinasi pemodelan yang diusulkan menggunakan *Lexicon Based* dan *Support Vector Machine*. Terdapat beberapa langkah alur sitem pada penelitian ini, mulai dari penarikan data latih. Data yang dilanjutkan tahap *preprocessing*, hingga dilakukannya pemodelan klasifikasi. Tahap-tahap yang dilakukan dalam penelitian ini meliputi:

#### a. Pengumpulan Data

Pada tahap ini dilakukan proses *scraping* konten data pada twitter akan diserap melalui API twitter dengan aplikasi pemrograman python, konten data yang diambil adalah tweet/kicauan masyarakat pengguna twitter yang berhubungan dengan kedua paslon presiden dan wakil presiden Indonesia pada Pemilu 2019.

#### b. Tahap Pra-pemrosesan

Pada tahap ini data yang diperoleh akan melewati beberapa proses seleksi diantaranya *case folding*, *tokenizing*, *filtering* maupun *stemming*. *Case folding* diperlukan untuk memastikan keseluruhan teks menjadi suatu bentuk standar yang umumnya terdiri dari keseluruhan huruf kecil atau *lowercase*. Proses *stop-word removal* juga dilakukan dengan cara mengeliminasi kata-kata yang tidak berpengaruh dalam klasifikasi, termasuk proses eliminasi URL. Selain itu proses juga akan menganalisis teks dan memastikan seluruh kata memiliki bentuk dasarnya, pada proses tokenisasi. Tahap *Stemming* adalah proses pencarian akar kata dengan mengeliminasi imbuhan pada kata, yang bertujuan untuk mengurangi jumlah token.

#### c. Metode Klasifikasi

Pada studi ini metode klasifikasi yang diusulkan adalah kombinasi dari *Support Vector Machine* dan *Lexicon Based*, hasil dari pre-proses teks akan dilakukan pelabelan dengan *Lexicon Based*. Pelabelan tersebut dilakukan untuk dataset yang akan diklasifikasi dengan *Support Vector Machine*, dengan pendekatan *Lexicon Based* untuk pelabelan dokumen maka pelabelan *dataset* untuk SVM tidak lagi dilakukan secara manual tetapi di proses oleh *Lexicon Based* dengan mengecek kata bersentimen positif, negatif, netral berdasarkan kamus atau lexicon pada kamus sentimen, dan menghitung frekuensi kemunculannya pada suatu dokumen teks.

#### d. Pengukuran Hasil Analysis

Pada hasil analisis ini akan menyajikan output dari klasifikasi *Support Vector Machine* berupa hasil aktual dan prediksi dari data training yang diolah oleh sistem pada setiap kelas juga hasil akurasi yang diperoleh dari data yang diolah.

#### e. Pengujian

Pengujian dilakukan terhadap tingkat akurasi klasifikasi yang dihasilkan oleh model. Pengujian terhadap tingkat akurasi klasifikasi yang dihasilkan oleh sistem bertujuan untuk mengetahui nilai *accuracy*, *precision* dan *recall*. Pengujian dilakukan dengan melibatkan data pelatihan dan data pengujian yang komposisinya berbeda. Pada setiap bagian pelatihan data akan dilakukan sepuluh kali pengujian dengan komposisi data yang berbeda-beda hingga mendapatkan nilai komposisi data yang paling akurat. Dengan demikian tingkat akurasi sistem dihitung berdasarkan nilai rata-rata dari setiap jumlah data pelatihan.

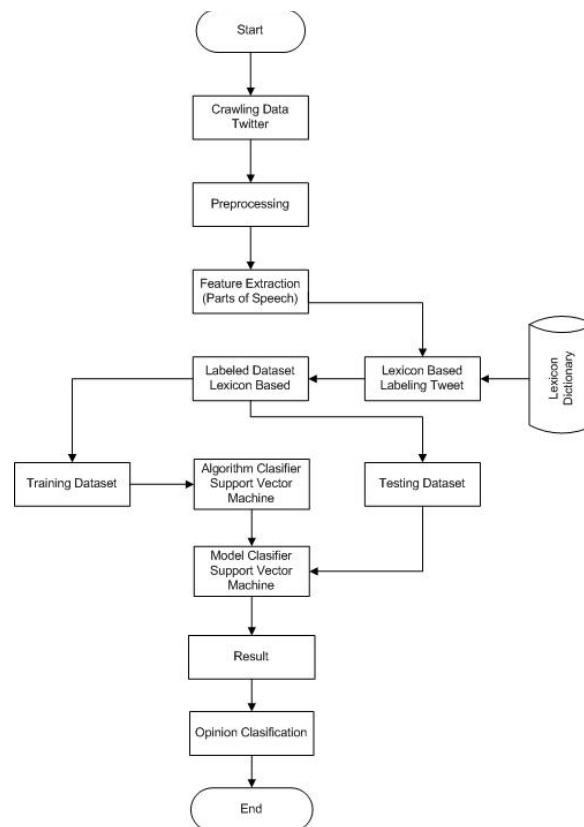
#### f. Metode Pemilihan Sampel

Pada studi ini, data diperoleh dengan proses *crawling* data twitter menggunakan aplikasi python, data yang diambil adalah *tweet* maupun *# (hashtag)* pada twitter yang berhubungan dengan keyword pada studi kasus penelitian ini yaitu konten yang mengandung nama pasangan calon presiden dan wakil presiden Indonesia untuk Pilpres 2019 yaitu “Prabowo Subianto-Sandiaga Uno” dan “Joko Widodo-Ma’ruf Amin”. Studi ini menganalisis 800 *corpus* Tweet. Berdasarkan studi terdahulu maka data dibagi secara imbang untuk setiap kelas [9], dengan pembagian corpus tentang Joko Widodo-Ma’ruf Amin sebanyak empat ratus Tweet, Prabowo Subianto-Sandiaga Uno juga sebanyak empat ratus Tweet.

#### g. Analisis Kombinasi Algoritma

Kombinasi algoritma dilakukan dengan melakukan proses *Lexicon Based* terlebih dahulu, *Lexicon Based* adalah pendekatan pada tingkatan kata, yaitu entitas yang diproses adalah kata dan tidak dapat mengidentifikasi kata-kata yang berupa singkatan, ekspresi atau bahasa percakapan. Selanjutnya data tweet yang telah diberikan label pada proses *Lexicon Based* yang digunakan sebagai data latih/training pada metode *Support Vector Machine*. Dengan begitu data training dapat ditransfer secara otomatis tanpa pelabelan secara manual. Selanjutnya data akan diproses klasifikasi dengan algoritma *Support Vector Machine* kedalam tiga kelas yaitu kelas positif, negatif dan netral.

Data yang diperoleh dari proses *Lexicon Based* akan melalui proses pembobotan terlebih dahulu dengan menggunakan TF-IDF karena seringkali suatu term muncul di sebagian besar dokumen yang mengakibatkan proses pencarian *term* unik akan terganggu. Dengan adanya IDF maka bobot suatu term akan berkurang jika kemunculannya tersebar pada dokumen yang dianalisis. Dengan pendekatan metode ini, diharapkan dapat meningkatkan hasil akurasi dari metode *Lexicon Based* juga dapat membantu proses pelabelan data pada SVM yang sebelumnya dilakukan secara manual menjadi otomatis dengan *Lexicon Based*.



Gambar 3. Flowchart Kombinasi Metode

#### h. Pengujian Algoritma

Berdasarkan proses klasifikasi yang telah dilakukan, selanjutnya maka model diuji dengan menggunakan metode *k-fold cross validation*, dengan nilai  $k=10$ . Pada tahap pengujian, analisis dilakukan dengan mencari nilai *accuracy*, *precision* dan *recall*.

Tabel 1. *Confussion Matrix*

Data Aktual	Data Prediksi		
	Positif	Negatif	Netral
Positif	TP	FP	PN
Negatif	FN	TNg	NgN
Netral	NP	NNg	TN

Sumber: [3]

Tabel diatas adalah hasil prediksi menggunakan *Support Vector Machine* yang diukur dari performa tiap-tiap kelas.

Keterangan:

- TP : Kelas kata terprediksi benar bernilai positif
- FP : Kelas kata positif terprediksi negatif
- PN : Kelas kata positif terprediksi netral
- FN : Kelas kata negatif terprediksi positif
- TNg : Kelas kata negatif terprediksi negatif
- NgN : Kelas kata negatif terprediksi netral
- NP : Kelas kata netral terprediksi sebagai kelas kata positif
- NNg : Kelas kata netral terprediksi negatif
- TN : Kelas kata terprediksi netral



#### i. Data Validasi

Pada tahap pengujian, studi ini menggunakan data dari pakar bidang sosial politik untuk mengetahui akurasi sentiment analysis. Dalam proses perhitungan *Recall* dan *Precision*, TP merupakan *True Positive*, yaitu jumlah dokumen yang dihasilkan model sesuai jumlah dokumen yang diberi label oleh pakar. Sementara itu, FP merupakan *False Positive* atau jumlah dokumen yang dianggap salah oleh pakar akan tetapi oleh model dianggap benar. FN merupakan *False Negative* yaitu jumlah dokumen yang dianggap benar oleh pakar namun oleh model dianggap sebagai nilai salah.

### IV. HASIL DAN PEMBAHASAN

#### a. Implementasi

Tahap implementasi awal yang dilakukan adalah melakukan *crawling* data dari media sosial twitter sesuai dengan studi kasus yang diangkat. Pada implementasi ini akan menjelaskan lebih detail proses demi proses yang dilakukan.

#### b. Pengumpulan Data Twitter

Pada studi ini, dilakukan penarikan data menggunakan metode *crawling* data tweet melalui API dengan bahasa pemrograman Python yang disimpan dalam bentuk JSON file. *Crawling* data dilakukan selama 4 bulan dimulai September-Desember 2018.

Tabel 2. Akuisisi Data Twitter

No	KOMENTAR
1	Jokowi pasti menang #2019JokowiKyaiMaruf #2019JokowiKyaiMaruf
2	RT @gabikinkembung: PASPAMPRES NYA PAK JOKOWI LUCU JUGA YA HAHAHAHA <a href="https://t.co/gwxd0LimeS">https://t.co/gwxd0LimeS</a>
3	RT @TsamaraDKI: Jadi oposisi yang kritis boleh. Tapi masa menutup mata dengan upaya pemerintah memulihkan Palu? Masa acara IMF pun mau dipo
4	Ayo Jokowi! #2019JokowiKyaiMaruf #2019JokowiKyaiMaruf
5	RT @PollingLagi: Jokowi Unggul Jauh dari Prabowo di Survei SMRC Anda Percaya ? Percaya ( Like ) Tidak Percaya ( Retweet )
6	@Kasrudd36802705 @hnurwahid Eksekusi Hukuman mati ada di era Jokowi. Era SBY tdk berani mengeksekusi, malahan di be <a href="https://t.co/1TPX8363ej">https://t.co/1TPX8363ej</a>
7	#2019JokowiKyaiMaruf Mari nikmati! Jokowi 2 Periode :) <a href="https://t.co/8o9yHGpZVE">https://t.co/8o9yHGpZVE</a> ichi Leonardo - 2019 JOKOWI 2 PERIO <a href="https://t.co/6K8xEbbs9X">https://t.co/6K8xEbbs9X</a>

Pada tabel 2 diatas dapat kita lihat bahwa data yang diperoleh masih utuh, dengan berbagai karakter dan url link sehingga data tersebut perlu diolah dari karakter-karakter yang tidak diperlukan sehingga data yang diproses hanya data yang sudah di bersih untuk di proses ketahap selanjutnya.

#### c. Preprocessing

Tahap pengolahan data asli yang sudah diakuisisi berupa data *text* dari twitter, tujuan dari preprosesing adalah untuk mengeliminasi *noise*, memperjelas fitur, dan mengubah data asli agar sesuai dengan kebutuhan pemodelan. Dalam *preprocessing* terdapat beberapa tahapan diantaranya adalah *Cleansing*, *Case Folding*, *Tokenizing*, *filtering* dan *Stemming*. Pada tahap *Cleansing* data yang diperoleh akan dibersihkan dari karakter-karakter seperti html, hastag, alamat situs (url), username twitter (@username) maupun tanda baca. Proses pembersihan karakter-karakter tersebut dari dokumen text dilakukan dengan algoritma/pseudocode berikut ini:

##### Algoritma 1: Twitter Data Cleansing

```

1) dataset = "file_tweet.csv";
2) Illegal_char = "http[^:space:]*", "#\S+", "@\S+";
3) do while not EOF;
4)     If (dataset contain illegal_char) then;
5)         string_replace(illegal_char with null);
6)     end if;
7)     next;
8) end do;
```

Pada tahap *case folding* dilakukan penyeragaman bentuk huruf, dari huruf kapital menjadi huruf kecil. Proses penyeragaman huruf/*case folding* dari dokumen text dilakukan dengan algoritma/*pseudocode* berikut ini:

*Algoritma 2: Case Folding*

```
1) data_clean = after_cleansing_document;  
2) do while not EOF;  
3)   string text=preg_replace ("/^[A-Za-z ]/", "", data_clean);  
  
4)   string_to_lowwer_case(text);  
5)   next;  
6) end do;
```

Pada proses selanjutnya dilakukan eliminasi pada *slang word* dan menghilangkan huruf yang berulang seperti “adaaa” menjadi “ada”. Proses normalisasi kalimat ini dilakukan dengan beberapa tahapan dengan algoritma/*pseudocode* berikut ini:

*Algoritma 3: Normalisasi Kalimat*

```
1) data_case_folding = after_case_folding_document;  
2) slang_word = "SlangWord.csv";  
3) do while not EOF;  
4)   If (data_case_folding contain slang_word) then;  
  
5)     string_replace (data_case_folding with phrase );  
  
6)   end if;  
7)   next;  
8) end do;
```

Tabel 5. Proses Normalisasi Kalimat

No	Bentuk teks
1	jokowi pasti menang
2	rt paspampres nya pak jokowi lucu juga ya hahahaha
3	rt jadi oposisi yang kritis boleh tapi masa menutup mata dengan upaya pemerintah memulihkan palu masa acara imf pun mau dipo
4	ayo Jokowi
5	rt jokowi unggul jauh dari prabowo di survei smrc anda percaya percaya suka tidak percaya retweet
6	eksekusi hukuman mati ada di era jokowi era sby tidak berani mengeksekusi malahan di
7	mari nikmati pokowi periode ichi leonardo jokowi

**d. Proses Kombinasi Algoritma**

Pada tahap kombinasi metode ini dilakukan dengan 2 tahapan yaitu tahap *Lexicon Based* yang hasil pelabelannya akan dijadikan sebagai data training dan proses klasifikasi menggunakan metode *Support Vector Machine*.

1. Pelabelan dengan *Lexicon Based*

Data yang telah melalui tahapan preprocessing akan diolah kedalam metode *Lexicon Based* dengan melakukan *Part-Of-Speech Tagging* yang merupakan proses pelabelan setiap kata sesuai dengan POS atau tag sesuai dengan kelas kata seperti kata keterangan, kata sifat, kata kerja, dan lainnya. Ada beberapa tahapan yang dilakukan dalam metode *Lexicon Based* diantaranya adalah ekstraksi kata kunci positif dan negatif, evaluasi negasi. Dari *dataset* sebanyak 800 *tweet* akan dilakukan pemberian label dengan metode *Lexicon Based* untuk tiga kelas yaitu kelas positif, negatif dan netral.



Tabel 6. Perubahan Ekstraksi Kata Kunci

Dokumen	Positif <i>Keyword</i>	Negasi <i>Keyword</i>	Negatif <i>Keyword</i>
D1	pasti,menang	-	-
D2	lucu	-	-
D3	pulih	-	kritis, oposisi
D4	pilih	-	-
D5	unggul, suka	Tidak	
D6	-	-	hukum, mati, berani
D7	nikmat	-	-

Jika kata positif, negatif di dalam sebuah kalimat sudah diketahui, maka dilakukan perhitungan bobot nilai dengan cara menjumlahkan nilai kata pada opini. Jika jumlah nilai opini pada kalimat berjumlah 1, maka kelas sentimen dari kalimat tersebut ditetapkan sebagai kelas positif, sebaliknya jika nilai opini dalam kalimat berjumlah 0, kelas sentimen dari kalimat tersebut ditetapkan sebagai kelas netral, dan jika nilai opini dalam kalimat bernilai -1, maka kelas sentimen dari kalimat tersebut ditetapkan sebagai kelas negatif.

Dari *dataset training* dan *testing* yang telah dilakukan pelabelan tersebut akan dilakukan proses klasifikasi dengan *Support Vector Machine* untuk mengetahui akurasi. Data hasil pelabelan dengan *Lexicon Based* diatas kemudian dilakukan proses validasi data untuk mengetahui seberapa besar ketepatan pelabelan *Lexicon Based* dengan hasil yang divalidasi oleh pakar. Dari hasil validasi data yang telah dilakukan pakar diperoleh persentasi ketepatan pemberian label sebesar 94,5% dengan penjabaran sebagai berikut:

Tabel 7. Hasil Validasi Data Oleh Pakar

Hasil Validasi Pakar	Hasil Pelabelan <i>Lexicon Based</i>		
	Positif	Negatif	Netral
Positif	306	1	12
Negatif	4	277	12
Netral	11	4	173

Dengan hasil melihat hasil validasi pada tabel 7 di atas maka akurasi dari pelabelan dengan *Lexicon Based* dapat diperoleh dengan persamaan sebagai berikut:

$$Accuracy = \frac{TP+TN+TN}{Total\ Tweet} = \frac{306+277+173}{800} = 94,5\%$$

Maka hasil akurasi berdasarkan validasi data yang dilakukan oleh pakar adalah 94,5%. Dengan mendapatkan nilai akurasi diatas 90% maka hasil pelabelan dengan metode *Lexicon Based* dapat digunakan sebagai metode pelabelan untuk *dataset* pada studi ini.

#### e. Pengujian dengan *Confusion Matrix*

Pada tahap ini akan dilakukan percobaan/pengujian terhadap kombinasi metode dengan menggunakan k-fold cross validation. Dibawah ini adalah tabel hasil dari data training yang akan di klasifikasi dengan svm, sebanyak 800 data. Dengan nilai 80% yaitu komposisi data training 80% dan data testing 20%. Pada komposisi data tersebut memiliki nilai *Precision*, *Recall* dan *F-Measure* sebagai berikut:

Tabel 8 Hasil Pengujian Terbaik

Data Aktual	Data Prediksi			<i>Precision</i>	<i>Recall</i>	<i>F-Measure</i>
	Positif	Negatif	Netral			
Positif	72	4	1	0,972	0,935	0,953
Negatif	0	58	4	0,920	0,935	0,972
Netral	2	1	18	0,857	0,782	0,818

Data Training 80% = 640 Data

Data *Testing* 20% = 160 data

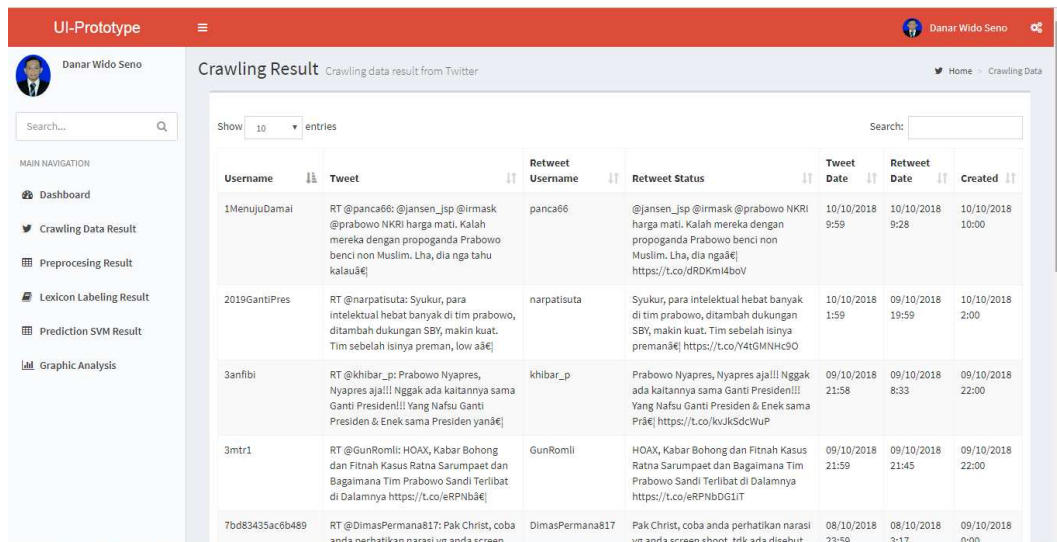
Nilai *Accuracy* = 92,5%

Nilai koefisiensi *Kappa* = 0,8762

Pada komposisi diatas *Precision* tertinggi dihasilkan pada prediksi kelas positif yaitu 92,5%.

#### f. Implementasi Prototipe

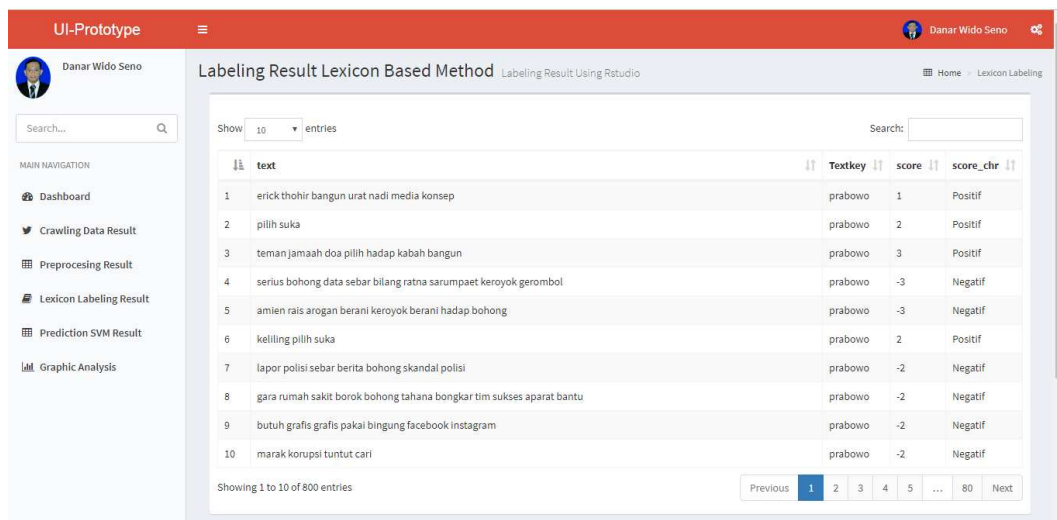
Pada tahap implementasi ini dilakukan dengan bahasa pemrograman R dan tools Rstudio serta PHP sebagai tampilan antar muka nya, agar dapat mempermudah membaca hasil data yang telah diolah oleh pemrograman R.



Username	Tweet	Retweet Username	Retweet Status	Tweet Date	Retweet Date	Created
1MenujuDamai	RT @panca66: @jansen_jsp @irmask @prabowo NKRI harga mati. Kalah mereka dengan propaganda Prabowo benci non Muslim. Lha, dia nga tahu kalaua	panca66	@jansen_jsp @irmask @prabowo NKRI harga mati. Kalah mereka dengan propaganda Prabowo benci non Muslim. Lha, dia nga	10/10/2018 9:59	10/10/2018 9:28	10/10/2018 10:00
2019GantiPres	RT @narpatissuta: Syukur, para intelektual hebat banyak di tim prabowo, ditambah dukungan SBY, makin kuat. Tim sebelah isinya preman	narpatissuta	Syukur, para intelektual hebat banyak di tim prabowo, ditambah dukungan SBY, makin kuat. Tim sebelah isinya preman	10/10/2018 1:59	09/10/2018 19:59	10/10/2018 2:00
3anfib	RT @khibar_p: Prabowo Nyapres, Nyapres aja!!! Nggak ada kaitannya sama Ganti Presiden!!! Yang Nafsu Ganti Presiden & Enek sama Presiden yan	khibar_p	Prabowo Nyapres, Nyapres aja!!! Nggak ada kaitannya sama Ganti Presiden!!! Yang Nafsu Ganti Presiden & Enek sama Pr	09/10/2018 21:58	09/10/2018 8:33	09/10/2018 22:00
3mtr1	RT @GunRomli: HOAX, Kabar Bohong dan Fitnah Kasus Ratna Sarumpaet dan Bagaimana Tim Prabowo Sandi Terlibat di Dalamnya	GunRomli	HOAX, Kabar Bohong dan Fitnah Kasus Ratna Sarumpaet dan Bagaimana Tim Prabowo Sandi Terlibat di Dalamnya	09/10/2018 21:59	09/10/2018 21:45	09/10/2018 22:00
7bd83435ac6b489	RT @DimasPermana817: Pak Christ, coba anda perhatikan narasi yg anda screen shoot, tdk ada disebut	DimasPermana817	Pak Christ, coba anda perhatikan narasi yg anda screen shoot, tdk ada disebut	08/10/2018 23:59	08/10/2018 3:17	09/10/2018 0:00

Gambar 4. Tampilan Hasil *Crawling* Data

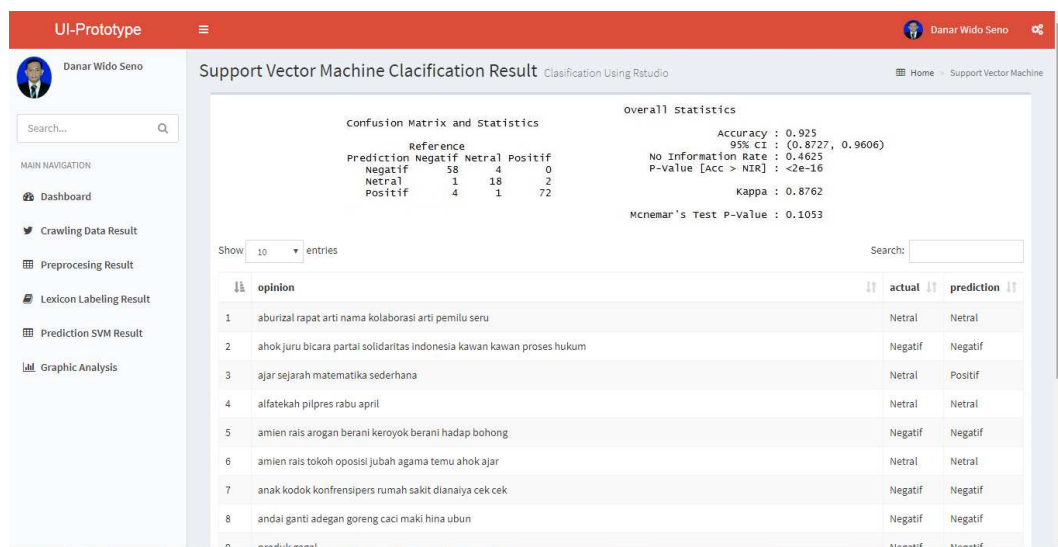
Pada Gambar 4 terlihat hasil data csv dari *crawling* yang dilakukan dengan pemrograman python yang kemudian ditampilkan dalam data *grid* prototipe.



text	Textkey	score	score_chr
1 erick thohir bangun urat nadi media konsep	prabowo	1	Positif
2 pilih suka	prabowo	2	Positif
3 teman jamaah doa pilih hadap kabah bangun	prabowo	3	Positif
4 serius bohong data sebar bilang ratna sarumpaet keroyok gerombol	prabowo	-3	Negatif
5 amien rais arogan berani keroyok berani hadap bohong	prabowo	-3	Negatif
6 keliling pilih suka	prabowo	2	Positif
7 lapor polisi sebar berita bohong skandal polisi	prabowo	-2	Negatif
8 gara rumah sakit borok bohong tahana bongkar tim sukses aparat bantu	prabowo	-2	Negatif
9 butuh grafis pakai bingung facebook instagram	prabowo	-2	Negatif
10 marak korupsi tuntutan cari	prabowo	-2	Negatif

Gambar 5. Tampilan Hasil Pelabelan Dengan *Lexicon Based*

Pada gambar 5 di atas dapat dilihat hasil skor *determine sentiment* dan opini dari setiap tweet dengan metode *Lexicon Based* yang di proses dengan Rstudio. Dengan begitu data yang telah di lakukan pelabelan ini akan diproses untuk klasifikasi dengan model *Support Vector Machine*.



Gambar 6. Tampilan Hasil Klasifikasi *Support Vector Machine*

Pada gambar 8 di atas terlihat hasil klasifikasi yang dilakukan oleh *Support Vector Machine* dengan Rstudio, dapat dilihat data *textual*, hasil prediksi SVM dan opini aktual yang diproses sebelumnya dengan metode *Lexicon Based*.

## V. KESIMPULAN

Berdasarkan hasil pada studi yang telah dilakukan, maka dapat diambil beberapa kesimpulan sebagai berikut:

1. Hasil pengujian terhadap *Accuracy*, *Precision* dan *Recall* pada kombinasi metode *Lexicon Based* dan *Support Vector Machine* mendapatkan akurasi terbaik 92,5% dengan komposisi data *training* sebanyak 80% dan data *testing* sebanyak 20% dengan nilai *Precision* pada setiap kelas diantara 85,7% - 97,2% dan nilai *Recall* setiap kelas diantara 78,2% - 93,5%. Dengan akurasi tersebut dapat terlihat bahwa akurasi yang diperoleh menjadi lebih tinggi dibandingkan dengan penelitian yang dilakukan dengan *Lexicon Based* tunggal maupun *Holistic Lexicon Based* pada multi kelas yang akurasinya hanya mencapai diangka 87% - 88% seperti yang telah dilakukan pada penelitian sebelumnya pada tinjauan studi penelitian.
2. Dengan menambahkan metode *Lexicon Based* pada sisi *preprocessing* dapat membuat proses pelabelan dataset *Support Vector Machine* tidak lagi dilakukan secara manual melainkan diproses oleh metode *Lexicon Based* dan *dictionary* pada *lexicon* dapat ditambahkan seiring dengan perkembangan konten data pada media sosial twitter. Dengan begitu proses transfer dataset pada *Support Vector Machine* akan menjadi lebih efisien dengan akurasi pelabelan mencapai 90% dengan catatan data *dictionary* pada setiap kelas sesuai dengan arti dan esensi yang ada pada Kamus Besar Bahasa Indonesia.

## DAFTAR PUSTAKA

- [1] Y. Setyanto, "Media Sosial sebagai Sarana Komunikasi Perusahaan dengan Media Media Sosial sebagai Sarana Komunikasi Perusahaan dengan Media," no. September, 2016.
- [2] G. Vinodhini, "Sentiment Analysis and Opinion Mining : A Survey," vol. 2, no. 6, 2012.
- [3] P. Nomleni, "Sentiment Analysis Menggunakan Support Vector Machine ( SVM )," 2015.
- [4] D. Maulina, R. Sagara, I. Komputer, and J. R. Utara, "Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document," vol. 2, no. 1, pp. 35–40, 2018.
- [5] N. Muchammad, S. Hadna, and P. I. Santosa, "Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen di Twitter," no. March, 2016.
- [6] Nurhuda, F. and Sihwi, S. W. (2014) 'Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier', Jurnal ITSMART, 2.
- [7] N. Vyrva, "Sentiment Analysis in Social Media," 2016.

- [8] F. M. Matulatuwa, E. Sedyono, and A. Iriani, "Text Mining Dengan Metode Lexicon Based Untuk Sentiment Analysis Pelayanan PT . POS Indonesia Melalui Media Sosial Twitter," vol. 2, no. 3, p. 5093, 2017.
- [9] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *Integer J.*, vol. 1, no. 1, pp. 32–41, 2017.
- [10] Imam Syafei dan Hendri Murfi, "Analisis Kinerja Kombinasi Metode Berbasis Lexicon dan Metode Berbasis Learning pada Analisis Sentimen Twitter" *FMIPA UI*, 2014.
- [11] Purba, I. D. C., Hidayati, H. and Gozali, A. A. (2014) 'Metode Holistic Lexicon-Based Untuk Analisis Sentiment Pada Dokumen Bahasa Indonesia (Studi Kasus: Tweets Mengenai Isu Sosial Kota Bandung)', Telkom University.