Analisis Sentimen Politik pada Twitter Menggunakan Metode Support Vector Machine (Studi Kasus : Pilpres 2019)

Alberi Meidharma Fadli Hulu[1], Kemas Muslim Lhaksmana[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1] alberihulu@students.telkomuniversity.ac.id, [2]kemasmuslim@telkomuniversity.ac.id

Abstrak

Salah satu media sosial yang digemari masyarakat adalah twitter. Banyak masyarakat yang mencurahkan pendapat atau pikiran mereka melalui tweet twitter, termasuk seperti pada masa sekarang ini masyarakat beramai-ramai memberikan tanggapan atau komentar melalui twitter mengenai peristiwa politik dan pemilihan presiden, komentar tersebut tidak hanya berisi hal-hal positif, tetapi juga ditemukan komentar yang bersifat negatif. Kelebihan twitter yaitu dapat diakses oleh semua orang dan kalangan sehingga dapat digunakan untuk memperkenalkan kandidat yang didukung, bertarung opini, berdebat, hingga menciptakan berita atau opini yang bersifat bohong untuk menyerang lawan politik. Tugas akhir ini bertujuan untuk mengetahui sentimen politik pengguna Twitter terhadap pemilihan presiden menggunakan metode *Support Vector Machine* dengan pembobotan TF-IDF. SVM dapat digunakan untuk mengklasifikasikan sentiman data atau kalimat yang diperoleh dari status twitter. Performansi sistem diukur berdasarkan *Confusion Matrix* dan akurasi. Nilai akurasi tertinggi yang didapatkan pada penelitian ini sebesar 62.88 % dengan TF-IDF menggunakan bentuk gabungan kata Unigram, Bigram dan Trigram. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah class pada input space.

Kata kunci : ***media sosial, twitter, text maining, support vector machine***

Abstract

One of the popular social media is Twitter. Many people pour out their opinions or thoughts through Twitter tweets, including as in the present time the community is giving a response or commenting via twitter about political events and presidential elections, the comments not only contain positive things, but also comments that are found negative. The advantages of Twitter are that it can be accessed by all people and circles so that it can be used to introduce candidates who are supported, fight opinions, argue, to create news or opinions that are false to attack political opponents. This final project aims to find out the political sentiments of Twitter users towards the presidential election using the Suppor Vector Machine method by weighting the TF-IDF. SVM can be used to classify sentiments of data or sentences obtained from twitter status. System performance is measured based on Confusion Matrix and accuracy. The highest accuracy value obtained in this study was 62.88% with TF-IDF using a combined form of the words Unigram, Bigram and Trigram. The SVM concept can be explained simply as an effort to find the best hyperplane that functions as a separator of two classes in input space.

Keywords: ***social media, twitter, text maining, support vector machine***

1. Pendahuluan

Pemilihan presiden dan wakil presiden Indonesia adalah pesta demokrasi 5 tahunan untuk seluruh lapisan masyarakat Indonesia, setiap warga negara Indonesa yang telah mencukupi umur atau telah memiliki ktp berhak untuk memberikan suara menentukan presiden dan wakil presiden Indonesia. Pemilihan presiden dan wakil presiden kali serentak dengan pemilihan legislatif yang akan digelar pada 17 April 2019 [1]. Dengan jumlah pemilih sekitar 192 juta pemilih [1] dapat dipastikan jika minat masyarakat membicarakan politik semakin meninggkat, dengan adanya dua kubu yang berbeda masyarakat tidak hanya membicarakan atau berdebat politik secara langsung diduni nyata tetapi juga meluas hingga ke sosial media, dimana orang yang saling tidak mengenal dapat saling beradu argumentasi.

Sosial media sudah menjadi kebutuhan masyarakat di era digital ini. Tanggapan atau opini seseorang dapat dieskpresikan dan disebarluaskan dengan cepat dan bebas. Dengan banyaknya masyarakat yang mengespresikan opini mereka di media sosial khususnya Twitter meningkatkan kemungkinan menelusuri sentimeten politik dari masyarakat pengguna media sosial. Banyaknya pengguna media sosial memungkinkan para peneliti menggunakan status seseorang menjadi sebuah data sentimen yang dapat diolah dan di analisis. Analisis sentimen yang dihasilkan dari pengolahan status atau *tweet* pengguna twitter akan menampilkan sentimen masyarakat pengguna twitter terhadapa pemilihan presiden dan wakil presiden.

Banyak metode yang digunakan untuk melakukan analisis sentimen seperti *Naive Bayes, Maximum Entropy* dan *Support Vector Machine*, dan lain-lain . Teknik *machine learning* menggunakan satu set pelatihan dan tes

ditetapkan untuk klasifikasi. Metode menggunakan machine learning dapat memberikan hasil yang lebih baik, namun klasifier supervised learning membutuhkan banyak data latih yang telah diberikan label. Tanpa diberikan data latih yang telah diberi label, supervised learning tidak dapat bekerja . Pemilihan metode Support Vector Machine dipilih karena memiliki kemampuan generalisasi dalam mengklasifikasikan suatu pattern dengan akurasi yang cukup tinggi [2].

2. Studi Terkait

Berdasarkan penelitian yang dilakukan pada tahun 2013 oleh Faisol Nurhada dan Tim pada jurnal yang berjudul : Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier, menunjukan akurasi hasil terbaik adalah sebesar 53%. Pada penelitian tersebut menggunakan 1400 data yang berasal dari twitter. Berdasarkan hasil penelitian lainnya yang dilakukan pada tahun 2018 oleh Muhamad Fajar hasil akurasi terbaik menggunakan metode SVM adalah sebesar 63.78%, pada penelitian tersebut menggunakan TF-IDF sebagai pembobotan dan bentuk penggabungan kata unigram, bigram dan trigram. Pada penelitian tersebut mengujikan 1000 data yang berasal dari twitter yang sudah melewati proses *preprocessing.*

2.1 Analisis sentiment

Sentiment Analysis atau *Opinion Mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis Sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif [3]. Analisis Sentimen dapat dibedakan berdasarkan sumber datanya, beberapa level yang sering digunakan dalam penelitian Analisis Sentimen adalah Analisis Sentimen pada level dokumen dan Analisis Sentimen pada level kalimat. Berdasarkan level sumber datanya Analisis Sentimen terbagi menjadi 2 kelompok besar yaitu *coarse-grained Sentiment Analysis* dan *fined-grained Sentiment Analysis*. Pada *Sentiment Analysis Coarse-grained*, Analisis Sentimen yang dilakukan adalah pada level dokumen. Secara garis besar fokus utama dari *Sentiment Analysis* jenis ini adalah menganggap seluruh isi dokumen sebagai sebuah sentiment positif atau sentiment negatif. *Fined-grained Sentiment Analysis* adalah *Sentiment Analysis* pada level kalimat. Fokus utama *fined-greined Sentiment Analysis* adalah menentukan sentimen pada setiap kalimat [3].

2.2 *Preprocesing*

Text Preprocessing merupakan tahapan dari proses awal terhadap teks untuk mempersiapkan teks menjadi data yang akan diolah lebih lanjut. Sebuah teks yang ada harus dipisahkan, hal ini dapat dilakukan dalam beberapa tingkatan yang berbeda. Suatu dokumen atau *tweet* dapat dipecah menjadi bab, sub-bab, paragraf, kalimat dan pada akhirnya menjadi potongan kata/token. Selain itu pada tahapan ini keberadaan digit angka, huruf kapital, atau karakter- karakter yang lainnya dihilangkan dan dirubah [15]. Beberapa pekerjaan yang umum dilakukan sebagai awal *preprocessing* adalah agregasi, penyampelan, pengurangan dimensi, pemilihan fitur, diskretisasi dan binerisasi, dan transformasi variabel. Maka dapat dikatakan *Preprocessing* adalah teknik maupun strategi yang bertujuan untuk membuat suatu data lebih mudah untuk dikelolah atau cocok untuk digunakan pada text mining yang tentunya bertujuan agar meningkatkan hasil dari analisis *text mining.*

2.3 Tf-Idf

Term Frequency dan *Invers Document Frequency* (Tf-Idf) merupakan metode untuk menghitung bobot setiap kata pada semua dokumen. Dalam sebuah komentar ataupun *tweet*, setiap term yang muncul akan dihitung jumlahnya. Semakin sering atau besar jumlah kemunculannya, maka nilai bobot yang dimiliki term tersebut semakin besar. Kuantitas term yang muncul disebut sebagai *Term Frequency (TF). Inverse Document Frequency (IDF)* ialah jumlah banyaknya kemunculan sebuah term pada sebuah komentar atau *tweet* yang tidak memiliki hubungan hingga menjadi kata yang tidak berbobot. Semakin sedikit sebuah dokumen memiliki term yang tidak umum, semakin besar nilai IDF nya. Rumus IDF ialah :

$$IDF = \log \frac{D}{Dfi} \quad (2.1)$$

Keterangan:

IDF = Nilai *inverse* dari DFi
D = Banyaknya *tweet* pada datasets
Dfi = Banyaknya *tweet* pada datasets yang mengandung kata ke-i

Nilai TF-IDF adalah:

$$TFIDF = TF * IDF \quad (2.2)$$

Keterangan:
TFIDF = Nilai bobot kata dalam sebuah datasets
TF = Nilai jumlah kemunculan sebuah kata pada *tweet*
IDF = Nila kata yang muncul dalam sebuah datasets

Nilai TFIDF yang bernilai 0 bermakna kata ke-i memiliki jumlah kemunculan satu kali pada satu tweet dalam datasets yang ada.

2.4 Media Sosial

Media sosial terdiri dari *internet-based* yang dapat dikembangkan berdasarkan ideologis dan teknologi web , media sosial memungkinkan penciptaan dan pertukaran *user-generated hinchcliffe* konten. Melalui media sosial , pengguna bisa mengunggah foto , video , musik , gambar , dan kitab kitab suci untuk berbagi ide , perasaan , pendapat , dan pengalaman dengan yang lain [4]. Pada dasarnya media sosial mempunyai beberapa jenis, yaitu: Pertama, proyek kolaborasi *website*, kedua, blog dan *microblog*, ketiga, konten atau isi, keempat, situs jejaring sosial, kelima, *virtual game world*, keenam, *virtual social world*.

2.5 Metode *Support Vector Machine*

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang *pattern recognition*. Sebagai salah satu metode *pattern recognition*, usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinya menempatkannya sebagai *state of the art* dalam pattern recognition, dan dewasa ini merupakan salah satu tema yang berkembang dengan pesat. SVM adalah metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization (SRM)* dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*. Tulisan ini membahas teori dasar SVM dan aplikasinya dalam bioinformatika, khususnya pada analisa ekspresi gen yang diperoleh dari analisa microarray [2].

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada input space [2]. Hyeperlane adalah generalisasi dari garis lurus atau bidang datar. Problem klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut. *Hyperplane* pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur margin hyperplane tersebut dan mencari titik maksimalnya. Margin adalah jarak antara *hyperplane* tersebut dengan pattern terdekat dari masing-masing *class*. *Pattern* yang paling dekat ini disebut sebagai *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM. Tiap data (*example*) dinotasikan sebagai xi $\in$ <D, i = 1,2,$\cdots$,N. N adalah banyaknya data. *Positive class* dinotasikan sebagai +1, dan *negative class* sebagai -1. Dengan demikian, tiap data dan label *class*-nya dinotasikan sebagai : yi $\in$ $\{-1,+1\}$. Diasumsikan bahwa kedua class tersebut dapat dipisahkan secara sempurna oleh hyperplane di D-dimensional *feature space*. *Hyperplane* tersebut didefenisikan sbb.

$$w.xi + b = 0 \quad (2.3)$$

Data xi yang tergolong ke dalam *negative class* adalah mereka yang memenuhi pertidaksamaan berikut :

$$w.xi + b \leq -1 \quad (2.4)$$

Adapun data xi yang tergolong ke dalam *positive class*, adalah mereka yang memenuhi pertidaksamaan :

$$w.xi + b \geq 1 \quad (2.5)$$

Keterangan :
w : vector bobot
xi : nilai atribut
b : skalar yang digunakan sebagai bias

Dalam memilih solusi untuk menyelesaikan suatu masalah, kelebihan dan kelemahan masing-masing metode harus diperhatikan. Selanjutnya metode yang tepat dipilih dengan memperhatikan karakteristik data yang diolah. Dalam hal SVM, walaupun berbagai studi telah menunjukkan kelebihan metode SVM dibandingkan metode konvensional lain, SVM juga memiliki berbagai kelemahan. Kelebihan SVM antara lain sbb [2]:

1. Generalisasi, didefinisikan sebagai kemampuan suatu metode *(SVM, neural network, dsb.)* untuk mengklasifikasikan suatu *pattern*, yang tidak termasuk data yang dipakai dalam fase pembelajaran metode itu. Vapnik menjelaskan bahwa *generalization error* dipengaruhi oleh dua faktor: *error* terhadap *training set*, dan satu faktor lagi yang dipengaruhi oleh dimensi *VC (Vapnik-Chervokinensis)*. Strategi pembelajaran pada *neural*

network dan umumnya metode *learning machine* difokuskan pada usaha untuk meminimimalkan *error* pada *training-set*. Strategi ini disebut *Empirical Risk Minimization (ERM).* Adapun SVM selain meminimalkan *error* pada *training-set*, juga meminimalkan faktor kedua. Strategi ini disebut *Structural Risk Minimization (SRM)*, dan dalam SVM diwujudkan dengan memilih *hyperplane* dengan margin terbesar. Berbagai studi empiris menunjukkan bahwa pendekatan SRM pada SVM memberikan error generalisasi yang lebih kecil daripada yang diperoleh dari strategi ERM pada neural network maupun metode yang lain.

2. *Curse of dimensionality*, didefinisikan sebagai masalah yang dihadapi suatu metode *pattern recognition* dalam mengestimasikan parameter (misalnya jumlah *hidden neuron* pada *neural network, stopping criteria* dalam proses pembelajaran dsb.) dikarenakan jumlah sampel data yang relatif sedikit dibandingkan dimensional ruang vektor data tersebut. Semakin tinggi dimensi dari ruang vektor informasi yang diolah, membawa konsekuensi dibutuhkannya jumlah data dalam proses pembelajaran. Pada kenyataannya seringkali terjadi, data yang diolah berjumlah terbatas, dan untuk mengumpulkan data yang lebih banyak tidak mungkin dilakukan karena kendala biaya dan kesulitan teknis. Dalam kondisi tersebut, jika metode itu "terpaksa" harus bekerja pada data yang berjumlah relatif sedikit dibandingkan dimensinya, akan membuat proses estimasi parameter metode menjadi sangat sulit. *Curse of dimensionality* sering dialami dalam aplikasi di bidang *biomedical engineering*, karena biasanya data biologi yang tersedia sangat terbatas, dan penyediaannya memerlukan biaya tinggi. Vapnik membuktikan bahwa tingkat generalisasi yang diperoleh oleh SVM tidak dipengaruhi oleh dimensi dari *input vector* . Hal ini merupakan alasan mengapa SVM merupakan salah satu metode yang tepat dipakai untuk memecahkan masalah berdimensi tinggi, dalam keterbatasan sampel data yang ada.
3. Landasan teori sebagai metode yang berbasis statistik, SVM memiliki landasan teori yang dapat dianalisa dengan jelas, dan tidak bersifat *black box*.
4. Feasibility SVM dapat diimplementasikan relatif mudah, karena proses penentuan support vector dapat dirumuskan dalam QP problem. Dengan demikian jika kita memiliki *library* untuk menyelesaikan QP problem, dengan sendirinya SVM dapat diimplementasikan dengan mudah. Selain itu dapat diselesaikan dengan metode sekuensial sebagaimana penjelasan sebelumnya.

Disamping kelebihannya, SVM memiliki kelemahan atau keterbatasan, antara lain:

1. Sulit dipakai dalam problem berskala besar. Skala besar dalam hal ini dimaksudkan dengan jumlah sample yang diolah.
2. SVM secara teoritik dikembangkan untuk problem klasifikasi dengan dua class. Dewasa ini SVM telah dimodifikasi agar dapat menyelesaikan masalah dengan class lebih dari dua, antara lain strategi *One versus rest* dan strategi *Tree Structure*. Namun demikian, masing-masing strategi ini memiliki kelemahan, sehingga dapat dikatakan penelitian dan pengembangan SVM pada *multiclass-problem* masih merupakan tema penelitian yang masih terbuka.

2.6 Confusion Matrix

Conffusion Matrix bekerja dengan cara mengolah data untuk membandingkan hasil prediksi dengan label sesungguhnya. *Conffusion Matrix* menggunakan sistem *Precision* dan *Recall. Precision* memanfaatkan ketepatan informasi yang dihasilkan oleh sistem dan *Recall* memanfaatkan tingkat keberhasilan sistem untuk mencari sebuah informasi. Berikut adalah tabel yang menjelaskan mengenai *Confusion Matrix*:

Tabel 1. Nilai *Confusion Matrix*

Actual/Classified	*Classified Positif*	*Classified Negatif*
Actual Positif	*True Positif*(TP)	*False Negatif*(FN)
Actual Negatif	*False Positif*(FP)	*True Negatif*(TN)

Berikut yang dapat diukur oleh *Confusion Matrix* :

a. Recall

Recall adalah perbandingan hasil klasifikasi dengan kelas sesungguhnya,

$$Recall = \frac{tp}{tp+fn} \tag{2.6}$$

b. Precision

Precision adalah perbandingan antara data yang terdeteksi benar dengan seluruh data prediksi pada suatu kelas,

$$Precision = \frac{tp}{tp+fp} \quad (2.7)$$

c. Accuray

Accuracy adalah perbandingan antara data yang terdeteksi benar dengan seluruh data hasil prediksi,

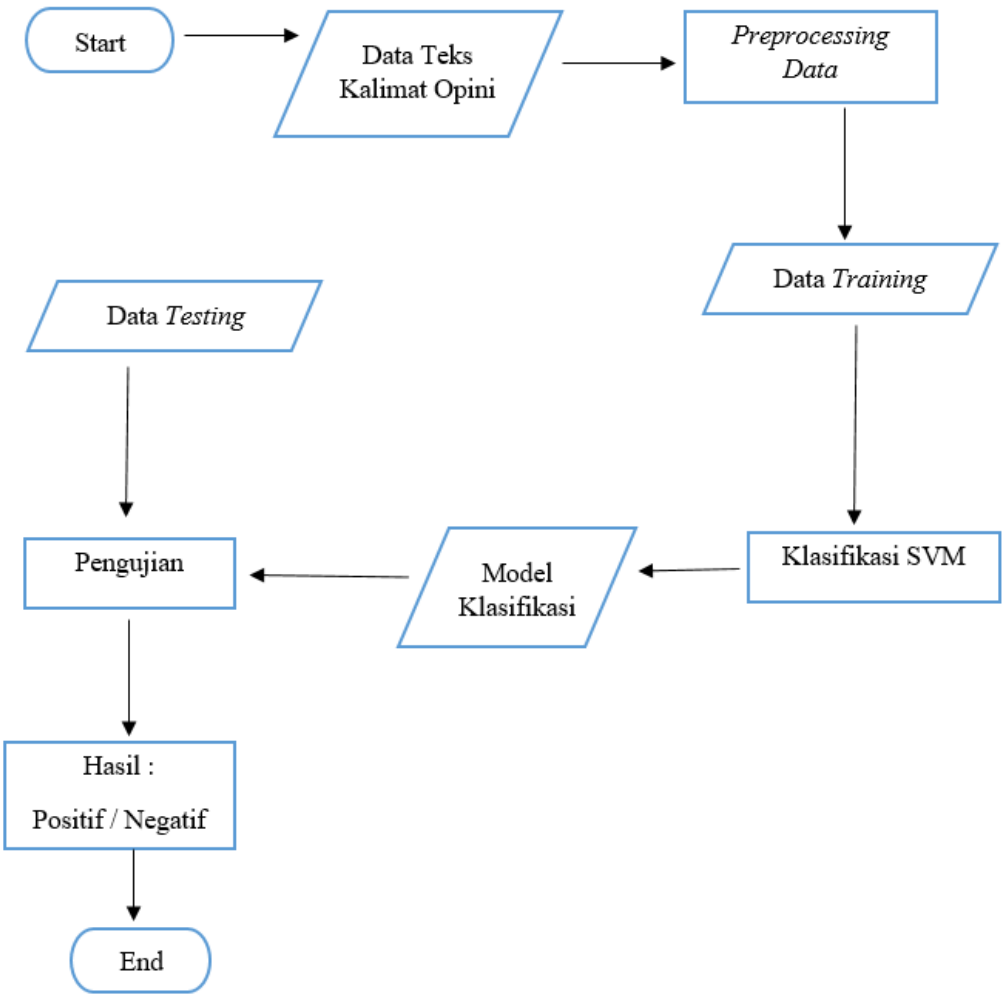
$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (2.8)$$

2. Sistem yang Dibangun

Pada bab ini akan dijabarkan secara lengkap mengenai sistem yang dibangun.

3.1 Gambaran Umum

Pada penelitian ini dibangun sebuah sistem yang dapat mengolah atau menganalisis sentimen dalam studi kasus Pilpres 2019 dengan menggunakan metode SVM. Data dibagi menjadi dua bagian, yaitu data training dan data testing. Baik data training maupun *data testing* merupakan hasil dari *tweet* yang diperoleh dengan cara *crawling*. Setiap data training telah melewati proses *preprocessing* dan pembobotan akan di train dengan menggunakan metode SVM untuk menghasilkan model dari data training tersebut. Setelah model yang didapat dari data training makan model tersebut akan digunakan sebagai model untuk melakukan testing. Sama halnya dengan *data training data testing* terlebih dahulu melewati tahapan preprocessing dan pembobotan. Berikut adalah gambar dari sistem yang akan dibangun.



Gambar 1. Gambaran Umum Sistem

a. Pengumpulan Data

Pengumpulan data status komentar atau pun *tweet* Twitter menggunakan *crawler* melalui API yang disediakan oleh Twitter dan menggunakan crawler yang disediakan netlytic.org . Jumlah data tweet yang berhasil dikumpulkan adalah 1000 data. Berikut adalah contoh *tweet* yang diperoleh:

Tabel 2. *Tweet* hasil dari *crawling*

NO	*Tweet*
1	mantap Pak terus majulah Indonesiaku Jokowi Presidenku @jokowi #Jokowi2Periode #JokowiLagi
2	Siap dukung Jokowi jempol @jokowi #Jokowi2Periode #JokowiAmin
3	kunci persatuan adalah Jokowi Maruf 2019 itu aja kok #JokowiAmin #salut #JokowiMembangunIndonesia
4	Tidak Ada Yang Terlihat Keren Seperti Capresku Prabowo Berwibawa sekali @prabowo @sandiuno @mulim_bersatu #2019gantipresiden #indonesiaadilmakmur
5	Dari dulu emang paling sayang dengan Prabowo Sandi sih @prabowo @sandiuno #2019gantipresiden #prabowopresiden #koalisiadilmakmur

b. *Preprocessing*

Tahap *Preprocessing* diperlukan untuk membersihkan data *Crawling* dari kata-kata yang tidak diperlukan seperti *hastag,mantion* dan URL. Menghilangkan kode atau symbol yang tidak terbaca sistem. Mengubah kata menjadi *non*-kapital.Menghilangkan kata *stopword*, yaitu kata yang tidak memiliki arti tertentu. Melakukan *stemming*, menemukan kata dasar sebuah kata, dengan menghilangkan semua imbuhan baik awalan maupun akhiran. bertujuan untuk memaksimalkan perhitungan dalam pemprosesan metode *SVM*. Serta memaksimalkan perhitungan yang nantinya akan dibuat. Berikut merupakan hasil *preprocessing* dari data yang digunakan :

Tabel 3. Hasil *Preprocessing Twitter*

NO	*Tweet* Asli	*Tweet Preprocessing*
1	mantap Pak terus majulah Indonesiaku Jokowi Presidenku @jokowi #Jokowi2Periode #JokowiLagi	mantap pak terus majulah indonesiaku jokowi presidenku jokowi jokowi2periode jokowilagi
2	Siap dukung Jokowi jempol @jokowi #Jokowi2Periode #JokowiAmin	siap dukung jokowi jempol jokowi jokowi2periode jokowiamin
3	kunci persatuan adalah Jokowi Maruf 2019 itu aja kok #JokowiAmin #salut #JokowiMembangunIndonesia	kunci persatuan adalah jokowi maruf 2019 itu saja kok jokowiamin salut jokowimembangunindonesia
4	Tidak Ada Yang Terlihat Keren Seperti Capresku Prabowo Berwibawa sekali @prabowo @sandiuno @mulim_bersatu #2019gantipresiden #indonesiaadilmakmur	tidak ada yang terlihat keren seperti capresku prabowo berwibawa sekali prabowo sandiuno mulim_bersatu 2019gantipresiden indonesiaadilmakmur
5	Dari dulu emang paling sayang dengan Prabowo Sandi sih @prabowo @sandiuno #2019gantipresiden #prabowopresiden #koalisiadilmakmur	dari dulu memang paling sayang dengan prabowo sandi sih prabowo sandiuno 2019gantipresiden prabowopresiden koalisiadilmakmur

c. Tokenisasi

Tokenisasi adalah memecah kalimat menurut jumlah katanya. Pada sistem ini tokenisasi dilakukan agar mendapatkan 3 buah bentuk kata. Bentuk kata tersebut adalah; Unigram, setiap token atau kata pada kalimat dibagi sejumlah satu kata. Bigram, setiap token atau kata pada kalimat dibagi sejumlah dua kata. Trigram, setiap token atau kata pada kalimat dibagi sejumlah tiga kata.

Tabel 4. Contoh Hasil Tokenisasi

NO	Bentuk Kata	Tokenisasi
1	Unigram	mantap, pak, terus, majulah, indonesiaku, jokowi, presidenku, jokowi, jokowi2periode, jokowilagi
2	Bigram	mantap pak, pak terus, terus majulah, majulah indonesiaku, indonesiaku jokowi, jokowi presidenku, presidenku jokowi, jokowi jokowi2periode, jokowi2periode jokowilagi
3	Trigram	mantap pak terus, pak terus majulah, terus majulah indonesiaku, majulah indonesiaku jokowi, indonesiaku jokowi presidenku, jokowi presidenku jokowi, presidenku jokowi jokowi2periode, jokowi jokowi2periode jokowilagi

d. Pembobotan Fitur

Pada tahapan ini menggunakan metode TF-IDF. Nilai TF diambil dari jumlah frekuensi kemunculan kata pada dokumen dan IDF diambil dari jumlah dokumen yang terdapat kata tersebut.

e. Training dengan SVM

Data yang sudah ditokenisasi dan melalui pembobotan fitur atau TF-IDF maka akan dilakukan training menggunakan metode SVM. Pada tahapan ini, data dibentuk agar sesuai dengan proses training pada SVM. Training pada SVM dilakukan sejumlah skenario yang diinginkan.

4. Hasil dan Analisis

Pada bab ini akan dijabarkan secara lengkap mengenai hasil beserta analisisnya.

4.1 Dataset dan Labelling

Dataset diambil menggunakan *crawler* yang dilakukan dalam kurung waktu selama enam bulan dimulai dari bulan Juni sampai dengan bulan November pada tahun 2018. Kata kunci yang digunakan pada *crawler* adalah sebagai berikut:

Tabel 5. Kata kunci untuk *crawler*

Kata Kunci		
Pilpres 2019	Pilpres	Jokowi
Prabowo	Sandiaga	Maaruf Amin
#2019gantipresiden	#2periode	PrabowoSandi
JokowiAmin	koalisi adil makmur	

Jumlah dataset yang digunakan sebesar 1000 *tweet*. *Tweet* tersebut kemudian dilabeli secara manual.

Tabel 6. Contoh Dataset dan Hasil Labeling

NO	*Tweet*	Label
1	kerja nyata presiden jokowi jokowi membangun Indonesia	Positif
2	prabowo dan sandiuno akan mengangkat martabat bangsa ini prabowo menang	Positif
3	si kodok pencitraan terus sedangkan listrik didaerah terpencil belum tersentuh	Negatif
4	golongan kampret mah bisanya bilang pendusta dan pembohong kalau tidak itu ya nyebar hoax	Negatif

Tabel 7 menampilkan jumlah label hasil dari *labelling* setiap *tweet*.

Tabel 7. Jumlah Data tiap-tiap label

NO	Label	Jumlah *Tweet*
1	Positif	561
2	Negatif	439

4.2 Skenario Pengujian

Skenario pengujian dibuat untuk menentukan hasil akurasi terbaik dalam sistem. Pada skenario ini digunakan parameter uji, yaitu, nilai TF-IDF, ada tidaknya kata pada *tweet* dan bentuk kata. Berikut adalah skenario yang akan digunakan:

Tabel 8. Skenario Pengujian

Parameter	Kode	Skenario
Pengujian mengggunakan TF-IDF	AL1	Pengujian bentuk kata Unigram dengan TF-IDF
	AL2	Pengujian bentuk kata Bigram dengan TF-IDF
	AL3	Pengujian bentuk kata Trigram dengan TF-IDF
	AL4	Pengujian bentuk kata Unigram + Bigram dengan TF-IDF
	AL5	Pengujian bentuk kata Unigram + Trigram dengan TF-IDF
	AL6	Pengujian bentuk kata Bigram + Trigram dengan TF-IDF
	AL7	Pengujian bentuk kata Unigram + Bigram + Trigram dengan TF-IDF
Pengujian tanpa TF-IDF	AL8	Pengujian bentuk kata Unigram tanpa nilai TF-IDF
	AL9	Pengujian bentuk kata Bigram tanpa nilai TF-IDF
	AL10	Pengujian bentuk kata Trigram tanpa nilai TF-IDF
	AL11	Pengujian bentuk kata Unigram + Bigram tanpa nilai TF-IDF
	AL12	Pengujian bentuk kata Unigram + Trigram tanpa nilai TF-IDF
	AL13	Pengujian bentuk kata Bigram + Trigram tanpa nilai TF-IDF
	AL14	Pengujian bentuk kata Unigram + Bigram + Trigram tanpa nilai TF-IDF

4.3 Hasil Pengujian

Sebelum dilakukan pengujian pada skenario yang telah dibuat, sistem terlebih dahulu melakukan pengujian berupa pembagian jumlah data menjadi data training dan data testing. Pengujian menggunakan skenario pada AL4. Berikut hasil dari pengujian pembagian jumlah data:

Pengujian dilakukan dengan menggunakan metode *Confusion Matrix* untuk mencari nilai akurasi dari setiap skenario yang telah dibuat. Tabel 9 menjeleaskan hasil dari pengujian yang didapat dari sistem analisis sentimen yang telah dibuat.

Tabel 9. Nilai akurasi hasil pengujian

Perbandingan Data	Akurasi Total
90:10	58.88%
85:15	57.68%
80:20	56.24%
70:30	47.68%
60:40	45.03%

Dari hasil pengujian diatas, maka digunakan perbandingan data sebesar 90% data training dan 10% data testing. Selanjutnya dilakukan pengujian berdasarkan skenario masing masing. Hasil pengujian untuk setiap skenario dapat dilihat pada tabel 10.

Tabel 10. Nilai Akurasi tiap skenario

Kode scenario	Akurasi
AL1	0.5888
AL2	0.5786
AL3	0.4252
AL4	0.5901
AL5	0.5742
AL6	0.4785
AL7	**0.6288**
AL8	0.5933
AL9	0.5777
AL10	0.4463
AL11	0.5978
AL12	0.5892
AL13	0.5873
AL14	**0.6278**

Dari hasil pengujian pada setiap skenario, didapatkan akurasi terbaik pada skenario AL7. Tabel 11 menampilkan *Confusion Matrix* dari skenario AL7.

Tabel 11. *Confusion Matrix* dari skenario terbaik

Kelas Sebenarnya	Kelas Prediksi	
	Positif	Negatif
Positif	18	5
Negatif	8	22

4.4 Analisis

Berdasarkan pengujian yang telah dilakukan dan merujuk pada tabel 9 maka dapat diketahui bahwa semakin kecil data training yang digunakan maka nilai akurasi juga akan semakin kecil. Hal ini disebabkan oleh semakin sedikitnya jumlah kata yang terbentuk pada proses yang mempengaruhi nilai TF-IDF. Dari penelitian tersebut dapat dianalisa bahwa nilai TF-IDF mempengaruhi tingkat akurasi dari sistem klasifikasi yang dibangun. Dapat dilihat dalam skenario AL7 Pengujian kata dengan TF-IDF Unigram + Bigram + Trigram yang mendapatkan nilai akurasi 62.88%. Berdasarkan informasi tersebut diketahui bahwa penggabungan bentuk kata dapat meningkatkan nilai besaran akurasi. Selain itu pengujian tanpa menggunakan TF_IDF juga mempengaruhi nilai akurasi yang didapatkan.

Pada tabel 11 *confusion matrix*, berdasarkan skenario yang digunakan didapatkan *True Positif* sebanyak 53 data dan *True Negatif* 37 data,dari situ dapat diketahui bahwa kalimat yang berhasil diprediksi dengan baik adalah tweet yang bersifat positif. Hal ini tersebut dikarenakan banyaknya kumpulan kata pada kelas positif yang terdapat pula pada kelas negatif.

5 Kesimpulan

Dalam penelitian ini sistem dibangun untuk melakukan klasifikasi sentimen politik pengguna twitter dengan menggunakan metode SVM dan pembobotan TF-IDF. Pengujian dijalankan dengan menggunakan beberapa skenario untuk menguji besaran pengaruh pembobotan kata menggunakan tokenisasi bentuk kata yaitu, unigram, bigram, dan trigram dalam proses melakukan klasifikasi.

Dari hasil pengujian yang telah dilaksanakan, maka didapatkan kesimpulan bahwa sentimen yang paling banyak muncul oleh pengguna Twitter mengenai Pilpres 2019 berdasarkan data yang digunakan adalah sentimen yang bernilai atau bersifat positif. Dengan menggunakan metode SVM dalam melakukan klasifikasi menghasilkan nilai akurasi sebesar 62.88%, nilai diperoleh melalui proses TF-IDF dan SVM menggunakan gabungan bentuk kata unigram, bigram, dan trigram secara bersamaan.

Nilai akurasi yang mencapai besaran 62.88% belum dapat dikatakan berhasil dengan sempurna karena jumlah data yang digunakan hanya sebesar 1000 data tweet. Faktor tersebut mengakibatkan sistem belum secara maksimal tepat memprediksi sentimen yang ada didalam data *tweet*. Akan tetapi dalam pengembangan kedepannya, dapat dilakukan penambahan jumlah data *tweet*, serta penambahan kata pada stopword dan penggunaan emotikon ataupun tanda baca lainnya.

Daftar Pustaka

[1] KPU, “KPU,” 2018 Available: https://infopemilu.kpu.go.id/

[2] Nugroho .A.S, “Pengantar Support Vector Machine” 8 Februari 2007

[3] Falahah dan Dyar Dwiki Adriadi Nur, “Pengembangan Aplikasi Sentiment Analysis Menggunakan Metode Naive Bayes”, 3 November 2015

[4] Linda S.L. Lai, “Content Analysis of Social Media :A Grounded Theory Approach”, 2015

[5] Mullen .T and Collier .N, “Sentiment analysis using support vector machines with diverse information sources’’.

[6] Abbasi .A, Rashidi .A.T, Maghrebi .M, Waller .S.T “Utilising Location Based Social Media in Travel Survey Methods: bringing Twitter data into the play”.

[7] Talentica Software, “Sentiment Analysis: Movie Reviews”, 2011

[8] Timmaraju .A and Khanna .V “Sentiment Analysis on Movie Reviews using Recursive and Recurrent Neural Network Architectures”.

[9] Detik.com, “KPU Tahapan Pemilu Serentak 2019.” Available: https://news.detik.com/berita/3483078/pileg-dan-pilpres-serentak-digelar-17-april-2019-ini-tahapannya

[10] Yessenov .Y and Misailovic .S “Sentiment Analysis of Movie Review Comments”, 20009

[11] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in First International Conference on Machine Learning, Rutgers University, New Brunswick: NJ, USA, 2003.

[12] Kapukaranov .B and Nakov .P “Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian”.

[13] Saraswati, N.W.S., Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machine untuk Sentimen Analysis, 2011

[14] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan “Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques”.

[15] Muktamar .B “Implementation of Data Mining With Naive Bayes Classifier to Supporting Marketing Strategy in Public Relations”, 2013