USULAN PENELITIAN S1

SENTIMENT ANALYSIS UNTUK MENGETAHUI ELEKTABILITAS CALON PRESIDEN DAN WAKIL PRESIDEN PADA PEMILIHAN UMUM 2019 DI INDONESIA

GAMA CANDRA TRI KARTIKA

15/378060/PA/16535

PROGRAM STUDI ILMU KOMPUTER
DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS GADJAH MADA
YOGYAKARTA

2018

HALAMAN PERSETUJUAN

USULAN PENELITIAN S1

SENTIMENT ANALYSIS UNTUK MENGETAHUI ELEKTABILITAS CALON PRESIDEN DAN WAKIL PRESIDEN PADA PEMILIHAN UMUM 2019 DI INDONESIA

Diusulkan oleh

GAMA CANDRA TRI KARTIKA
15/378060/PA/16535

Telah disetujui oleh Tim Penguji
Pada tanggal 5 Desember 2018

Susunan Tim Penguji,

Drs. Sri Mulyana, M.Kom	Dr. Agus Sihabudin, S.Si., M.Kom.
Pembimbing	Ketua Penguji
	Drs. Edi WInarko, M.Sc., Ph.D.
	Penguji

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Media sosial adalah teknologi yang menggunakan komputer secara interaktif yang dapat memfasilitasi untuk membuat serta berbagi informasi, ide, opini, dan bentuk ekspresi lain melalui komunitas dan jaringan virtual seperti internet. Menurut survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), Pada tahun 2017, jumlah pengguna internet di Indonesia mencapai 143,26 juta jiwa. Angka tersebut meningkat 7.95 persen dibandingkan pada tahun sebelumnya, yang mencapai 132,7 juta jiwa (Yuniarni, 2018).

Indonesia merupakan negara demokrasi yang melaksanakan pemilihan umum (Pemilu) setiap 5 tahun sekali untuk memilih pemimpin beserta perangkat legislatif lainnya. Sampai saat ini, pemilu masih digunakan sebagai inti dari proses demokrasi di Indonesia. Pada pemilu 2019 nanti, sudah ditentukan kedua calon presiden dan wakil presiden yaitu calon pertama Ir. H. Joko Widodo dengan wakilnya Prof. Dr. K. H. Ma'ruf Amin dan calon kedua Letnan Jenderal (Purn.) H. Prabowo Subianto Djojohadikusumo dengan pasangannya H. Sandiaga Salahuddin Uno, B.B.A., M.B.A.

Media sosial seperti Twitter sudah dikenal di berbagai pengguna untuk penggunaanya sebagai penyampaian opini secara terbuka atas penyampaian pendapatnya pada suatu tokoh di media sosial. Hal ini dapat memberikan dampak yang besar bagi individu karena memuat berbagai opini dan pandangan terhadap seseorang. Berdasarkan usia, sebanyak 16,68 persen pengguna media sosial berusia 13-18 tahun dan 49,52 persen berusia 19-34 tahun. Hal ini menunjukkan cukup tingginya pemilih baru dalam menggunakan media sosial (Yuniarni, 2018)..

Pengaruh media sosial sangat besar dalam dunia politik. Sejak tahun 2008, para pasangan calon presiden Amerika Serikat telah menggunakan media sosial seperti Facebook dan Twitter untuk menggalang dukungan. Di Indonesia, pada pemilu presiden RI tahun 2014 lalu, tim kampanye pasangan calon dan juga para pendukungnya dengan gencar menggunakan media sosial dengan mengunggah

beragam video, foto atau pun status seputar pilpres melalui media sosial. Salah satu calon presiden, Joko Widodo, bahkan sudah menggunakan media sosial sebagai media kampanye ketika mencalonkan diri sebagai gubernur DKI Jakarta pada 2012 lalu. Hal ini pun tetap diikuti pada masa kampanye pemilu 2019 dengan tagar-tagar unik di Twitter yang menunjukkan dukungan pada masing-masing calon.

Analisis sentimen merupakan ilmu yang berguna untuk menganalisis pendapat seseorang, sentiment seseorang, evaluasi seseorang, sikap seseorang dan emosi seseorang ke dalam bahasa tertulis. Teknik sentimen dapat mendukung beberapa skenario seperti sentimen yang diskrit yang terdiri 1 dan 0 atau sentimen yang bernilai kontinyu.

Elektabilitas adalah tolak ukur tingkat keterpilihan dari calon tersebut berdasarkan kriteria pilihan saat akan maju kedalam pemilihan umum. Elektabilitas seseorang dipengaruhi oleh opini pendukung calon presiden bagaimana prospek nominasi kedepanya. Dengan banyak opini yang positif terhadap calon presiden, semakin banyak massa yang mendukung dan meningkatkan elektabilitas mereka (Abramowitz, 1989).

Telah banyak metode statistik yang dilakukan untuk memprediksi elektabilitas tokoh ataupun partai politik. Misalnya dengan survey, namun hasil survei yang dihasilkan oleh lembaga survey terkadang tidak sesuai dengan kenyataan dan seringkali digunakan untuk mengarahkan opini sehingga tidak netral oleh salah satu kandidat (Lestari dkk, 2017). Terdapat juga quick count, namun quick count membutuhkan waktu yang lama dan dilaksanakan setelah pemilihan umum. Pada penelitian yang lain, metode analisis sentimen dengan berbagai algoritma telah diusulkan untuk mengetahui dan mengevaluasi elektabilitas tokoh atau partai politik yang dilaksanakan sebelum pemilu (Lestari dkk, 2017).

Preprocessing perlu dilakukan untuk mengetahui didalam penelitian apakah memang baik dilakukan semua atau tidak dalam penelitian analisis sentimen, sehingga terlihat perbandingan akurasi dari data yang sudah dinormalisasi dan stemming (Saputra dkk, 2015). Dengan melihat hasil akurasi yang didapat, dapat disimpulkan apakah penggunaan *preprocessing* tersebut harus dilakukan atau tidak guna mendapatkan hasil akurasi yang optimal.

Performa model klasifikasi menjadi bagian penting dalam proses klasifikasi.Hal ini menunjukan seberapa akurat sistem dapat mengklasifikasikan data denganbenar. Salah satu metode untuk meningkatkan akurasi dengan seleksi fitur. Seleksifitur adalah proses mereduksi fitur-fitur yang dianggap tidak relevan dalam prosesklasifikasi yang akan menimbulkan overfitting. Jika seleksi fitur TF-IDFmemperhitungkan jumlah kemunculan fitur saja, seleksi fitur chi squaremenggunakan metode statistika untuk mengukur independensi sebuah term dengankategorinya, tidak sebatas kemunculan fitur saja. Hal ini membuat performa model chi square lebih baik dari TF-IDF (Lestari dkk, 2017).

Metode klasifikasi yang dipakai dalam penelitian ini adalah *Support Vector Machines* (SVM). Alasan penggunaan metode klasifikasi SVM adalah karena pada dataset yang berisi ulasan yang lebih panjang, dibandingkan dengan kinerja *Multinomial Naïve Bayes* (MNB) yang sangat baik pada cuplikan data, banyak asumsi buruk tentang MNB menjadi lebih merusak bagi dokumen yang lebih panjang ini. SVM jauh lebih kuat daripada MNB untuk analisis sentimen dengan panjang menyeluruh, tetapi masih lebih buruk daripada beberapa hasil yang dipublikasikan lainnya (Wang dan Manning, 2012).

Karena pemilih yang menggunakan Twitter dan berita memiliki karakteristik kritis, mandiri, independen, rasional dan pro perubahan (Sukendar, 2017) penelitian ini mengusulkan topik analisis sentimen pada data Twitter dan berita untuk mengetahui elektabilitas tokoh politik dalam periode waktu tertentu berdasarkan sentimen (positif dan negatif) masing masing tokoh politik.

1.2 Rumusan Masalah

Data dari media sosial seperti Twitter merupakan data yang objektif untuk diolah dikarenakan penggunanya yang sangat aktif dalam menyampaikan pendapatnya di ketiga media sosial tersebut. Namun metode seperti sentimen analisis belum banyak diterapkan dalam domain Bahasa Indonesia. Metode survey yang dilakukan oleh lembaga survey memiliki kelemahan seperti membutuhkan proses yang lama, biaya yang mahal dan hasilnya terkadang tidak objektif. Sementara, metode ilmiah seperti analisis sentimen masih belum dimanfaatkan

lembaga survey di Indonesia. Oleh karena itu, pada penelitian ini akan digunakan metode analisis sentimen untuk mengukur elektabilitas kedua calon dengan harapan dapat diperoleh hasil dengan cepat dan lebih objektif.

1.3 Batasan Masalah

Batasan masalah dalam penilitian ini adalah sebagai berikut:

1. Penelitian ini hanya membahas tentang Pemilihan Umum 2019 dan tokoh-tokoh seperti Joko Widodo, Ma'ruf Amin, Prabowo Subianto, dan Sandiaga Uno.
2. Dataset yang digunakan berasal dari Twitter dan Line-Today.
3. Bentuk dataset hanya menampilkan tanggal terbit, sumber artikel dan hasil komentar saja.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengetahui elektabilitas kedua calon dengan pemilihan kombinasi yang terbaik dari proses normalisasi data dan stemming, terhadap akurasi analisis sentimen dengan metode SVM.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Dapat digunakan dalam survei untuk mengetahui elektabilitas sebelum mencalonkan diri dalam pemilihan umum atau daerah.
2. Dapat digunakan sebagai bahan evaluasi dalam proses kampanye.
3. Dapat mengetahui faktor yang paling berkorelasi dalam nilai elektabilitas.

BAB II
TINJAUAN PUSTAKA

Lestari, dkk (2017) meneliti analisis sentimen tentang opini pilkada DKI 2017 pada dokumen Twitter. Dokumen tersebut terkadang memuat unsur non-tekstual seperti adanya emoji. Emoji biasanya digunakan untuk mengungkapkan perasaan seseorang. Algoritma yang digunakan adalah Naïve Bayes dengan melakukan pembobotan non-tekstual (*emoji*) dan pembobotan tekstual. Hasil dari penelitian ini berupa sentimen positif dan negatif yang sudah dinormalisasi dengan *Min-Max Normalization* dengan nilai akurasi yang didapat yaitu 68,52% untuk pembobotan tekstual dan 75,93% untuk pembobotan non-tekstual. Dengan ini diketahui bahwa pembobotan non-tekstual berpengaruh terhadap akurasi dan pengklasifikasian yang didapat.

Prasetyo (2014) meneliti sebuah metode untuk memprediksi hasil pemilu Presiden RI 2014 dengan melakukan analisis pada data Twitter. Analisis dilakukan dengan membandingkan nilai *Mean Absolute Error* (MAE) dari sembilan faktor yang berbeda terhadap hasil KPU. Sembilan faktor tersebut adalah *tweet count*, durasi data, seleksi kata kunci, *user count*, penyaringan data, berat populasi, informasi kelamin, analisis sentimen, dan jumlah pengguna. Metode ini mampu mencapai MAE terendah, yakni 0,62% dengan menerapkan analisis sentimen pada masing-masing tweet. Sentimen masing-masing tweet diklasifikasikan dengan menghitung nilai *polaritas* sentimen dengan menggunakan metode *Sentiment Lexicon*. Perhitungan prediksi hasil pilkada dilakukan dengan membandingkan jumlah tweet positif untuk masing-masing pasangan calon, kemudian dihitung persentasenya. Hasil dari penelitain ini jauh lebih baik bila dibandingkan dengan metode prediksi konvensional yang dilakukan oleh beberapa lembaga survei independen.

Wicaksono, dkk (2016) meneliti metode untuk memprediksi hasil pemilu Presiden Amerika Serikat 2016 dengan melakukan analisis sentimen pada media sosial. Media sosial yang digunakan adalah Twitter dengan tweet berbahasa Inggris.

Ada tiga metode klasifikasi yang digunakan dalam penelitian ini, Binary Multinomial Naive Bayes, SentiWordNet, dan AFINN-11. Setelah itu, ketiga metode tersebut dibandingkan akurasinya. AFINN-111 dan Binary Multinomial Naive Bayes keduanya memberikan skor yang bagus. Namun, Binary Multinomial Naive Bayes dipilih untuk digunakan karena menghasilkan nilai F1- score yang lebih tinggi daripada AFINN-111.

Saputra, dkk (2015) meneliti analisis sentimen dari Presiden Indonesia periode 2014-2019 Joko Widodo. Metode yang digunakan dalam pengambilan data menggunakan *Search Techniques*. Setelah itu proses *preprocessing* dilakukan dengan cara s*temming* menggunakan *library* Sastrawi, tokenisasi N-gram, *stopword removal* dan mempertahankan *emoticon*. Metode klasifikasi yang digunakan hanya menggunakan SVM dengan akurasi yang terbaik dalam penelitian ini dengan dilakukan normalisasi dan stemming pada data sebesar 89,2655% menggunakan metode SVM, dan kemudian data yang dinormalisasi saja sebesar 88,7006% menggunakan metode SVM.

Razzaq, dkk (2014) meneliti analisis sentimen dari pemilihan umum Pakistan 2013. Media sosial yang digunakan adalah Twitter dengan kata kunci partai politik dan tokoh politik di Pakistan. Metode yang digunakan untuk preprocessing terdiri dari N-gram filter, *Laplace Smoothing*, *Porter Stemmer* dan TF-IDF untuk mencari kata yang paling relevant dengan dokumen. Lalu menggunakan perangkat Rainbow dan Weka, digunakan algoritma Naive Bayes, K-Nearest Neighbour, dan Prind untuk Rainbow lalu Random Forest, Support Vector Machine , dan Naive Bayes untuk Weka.

Alashri, dkk (2016) meneliti analisis sentimen dari pemilihan umum 2016 Amerika Serikat. Media sosial yang digunakan adalah Facebook kata kunci nama kandidat yang ditentukan. Metode dalam pengelompokan topik menggunakan *Latent Dirichlet Allocation* (LDA). Lalu metode untuk mengetahui korelasi aktifitas online dan offline menggunakan *Oblique Cumulative Curves* untuk menunjukkan tren yang lebih bersih dan jelas dibandingkan tren yang mentah. Wavelet Transform untuk noise pada data dan secara tepat mengidentifikasi tren peristiwa utama dalam kurva yang dihasilkan dari sentimen komentar.

Virgo (2018) juga melakukan penelitian analisis sentimen untuk memprediksi hasil Pilkada DKI 2017 oleh pasangan Anies-Sandi dan Ahok-Djarot. Algoritma yang digunakan adalah *Multinomial Naïve Bayes* dengan deteksi *Buzzer* tanpa seleksi fitur. Hasil dari penelitian ini , sistem dapat mengklasifikasikan sentimen untuk pasangan Ahok-Djarot dengan akurasi sebesar 77,28% dan pasangan Anies-Sandi dengan akurasi sebesar 79,70%.

Metode Naïve Bayes juga dilakukan oleh Hidayatullah dan Azhari (2014) untuk melakukan analisis sentimen dan klasifikasi tokoh publik pada Twitter. Tokoh publik yang dipilih adalah tokoh publik yang dianggap layak dan memiliki kemampuan untuk menjadi pemimpin. Naïve Bayes dikombinasikan dengan fitur sehingga dapat mendeteksi negasi dan menggunakan pembobotan *Term Frequency* dan TF-IDF. Selain metode Naïve Bayes juga digunakan metode *Support Vector Machine* (SVM) . Hasil klasifikasi berupa sentimen positif dan negatif dengan kategori tokoh politik berdasar kapabilitas, integritas, dan akseptabilitas tokoh tersebut. Dari proses pengklasifikasian sentimen dan kategori tokoh politik memang SVM lebih unggul dari Naïve Bayes.

Suryotomo (2018) melakukan penelitian dengan melakukan analisis sentimen menggunakan data tweet dan berita dari masing-masing tokoh politik untuk mengetahui elektabilitasnya menggunakan metode Multinomial Naive Bayes. Tokoh politik yang digunakan dalam penelitian adalah 10 tokoh politik yang dianggap populer di Indonesia. Dataset yang digunakan berjumlah 16.523 data *training* dan 6.550 data *testing*. Data tweet didapatkan menggunakan tool tweetcatcher dan berita didapatkan dari 3 situs berita di Indonesia yaitu tribunnews.com, tempo.co, dan viva.co.id menggunakan *tools scrapper* dalam kurun waktu 17 November 2016 sampai 1 November 2017. Setelah data terkumpul, dilakukan tahap preprocessing dan filtering. Lalu dilakukan seleksi top-n kata fitur menggunakan metode *chi square* dan TF-IDF. Selanjutnya adalah pembentukan model klasifikasi dan proses testing dengan membandingkan hasil elektabilitas tiap tokoh politik tanpa seleksi fitur dan dengan seleksi fitur chi square dan TF-IDF. Hasil penelitian ini menunjukan bahwa nilai performa model menggunakan metode seleksi fitur chi square lebih tinggi dengan rata-rata nilai akurasi 85,24%, presisi

88,84%, *recall* 91,65% dan *f-measure* 90,17% dibandingkan dengan menggunakan metode seleksi fitur TF-IDF dengan rata-rata nilai akurasi 78,11%, presisi 87,41%, *recall* 87,79% dan *f-measure* 87,54% serta jika dibandingkan tanpa seleksi fitur dengan nilai rata-rata akurasi 74,69% , presisi 87,40%, *recall* 84,88% dan *f-measure* 84,72%.

Hakimi (2018) juga meneliti tentang pemilihan kepala daerah diadakan serentak di sebagian besar daerah di Indonesia. Salah satu daerah tersebut adalah Jawa Timur. Dalam pra-pelaksanaan pemilihan Kepala Daerah Jawa Timur, terdapat berbagai opini masyarakat yang bersentimen positif dan negatif pada twitter. Opini tersebut dapat digunakan sebagai parameter untuk mengukur kekuatan masing-masing calon kepala daerah. Tujuan penelitian ini yaitu untuk mengetahui kecenderungan opini masyarakat tentang pemilihan Kepala Daerah Jawa Timur pada twitter. Dalam penelitian ini dilakukan beberapa tahapan yaitu crawling data twitter, pelabelan data, penghapusan data yang tidak dibutuhkan (data *outlier* dan netral), pre-proses data teks, pembuatan sistem klasifikasi dengan Naïve Bayes Classifier, dan uji coba sistem pada data twitter yang lebih banyak. Metode pembobotan kata yang digunakan yaitu *Term Frequency* (TF) dan *Term Frequency-Inverse Document Frequency* (TF-IDF). Diantara kedua metode tersebut, pembobotan kata TF mempunyai hasil yang lebih baik. Dari penelitian ini diperoleh hasil performa sistem untuk masing-masing kandidat. Untuk data calon gubernur nomor urut satu didapat hasil akurasi, presisi, *recall*, dan *f-measure* berturut-turut sebagai berikut 98,99%, 93,44%, 97,78%, dan 95,56%. Selanjutnya untuk data calon kepala daerah dengan nomor urut dua diperoleh hasil akurasi, presisi, *recall*, dan *f-measure* berturut-turut sebagai berikut: 98,95%, 97,78%, 98,55%, dan 98,17%. Berdasarkan data yang diperoleh dari twitter, masyarakat Jawa Timur cenderung lebih memilih calon kepala daerah dengan nomor urut satu yaitu Khofifah Indar Parawansa. Untuk perbandingan penelitian ditampilkan pada Tabel 2.1.

Tabel 2.1 Perbandingan penelitian yang sudah ada

No	Peneliti	Topik	Metode	Perbedaan

	Lestari dkk (2017)	Analisis sentimen untuk mengetahui opini masyarakat terhadap Pilkada DKI 2017	Sentimen Analysis Naïve Bayes Pembobotan emoticon	Bahasa dari dataset Metode klasifikasi Seleksi fitur menggunakan Chi-Square Pembobotan TF-IDF
	Prasetyo (2014)	Tweet-based election prediction	Membandingkan nilai MAE pada sembilan faktor Klasifikasi sentiment menggunakan Sentiment lexicon faktor salah satu analisis	Dataset Metode pengambilan nilai akurasi Perhitungan nilai sentimen
	Wicaksono dkk (2016)	A proposed method for predicting US presidential election by analyzing sentiment in social media	Membandingkan performa algoritma Binarized Multinomial Naïve Bayes, SentiWordNet, dan AFINN-11 untuk klasifikasi sentimen	Dataset Metode klasifikasi Pertimbangan dalam pemilihan skor dalam penilaian akurasi
	Saputra dkk (2015)	Analisis sentimen data	Metode pengambilan data	Dataset

		presiden Jokowi dengan preprocessing normalisasi dan stemming menggunakan metode Naïve Bayes dan SVM	menggunakan search techniques Preprocessing yang memanfaatkan library Sastrawi Metode klasifikasi SVM	Metode pengambilan data Stemming dalam bahasa Indonesia
	Razzaq dkk (2014)	Prediction and Anlysis of Pakistan election 2013 based on Sentiment Analysis	Preprocessing menggunakan N-Gran filter, Laplace Smoothing, Prter stemmer dan TF-IDF Algoritma yang digunakan Naïve Bayes	Dataset Langkah Preprocessing
	Alashri dkk (2016)	An analysis sentiment on facebook during the 2016 U.S. Presidential election	Klasifikasi kandidat berdasarkan topik yang ditentukan Mempelajari korelasi menggunakan wavelet transform	Dataset Metode klasifikasi menggunakan topic modelling
	Virgo (2018)	Analisis sentiment dan klasifikasi buzzer untuk prediksi	Klasifikasi menggunakan Naïve Bayes Deteksi Buzzer	Dataset Seleksi fitur Chi-Square dan TF-IDF

		Pilkada DKI 2017		Elektabilitas tokoh politik
	Hidayatullah dan Azhari (2014)	Analisis sentiment dan klasfikasi tokoh publik berdasarkan data di twitter	Metode klasifikasi menggunakan Naïve Bayes dan SVM Kategori klasifikasi pada tokoh publik	Dataset Metode klasifikasi Elektabilitas tokoh politik Metode seleksi fitur Chi-Square
	Suryotomo (2018)	Analisis sentimen untuk mengetahui elektabilitas tokoh politik menggunakan metode multinomial naïve bayes	Klasifikasi menggunakan metode Naïve Bayes Metode perhitungan elektabilitas	Dataset Elektabilitas tokoh politik Metode seleksi fitur Chi-Square dan TF-IDF
	Hakimi (2018)	Sistem analisis sentimen publik tentang opini pemilihan kepala daerah jawa timur 2018 pada dokumen twitter menggunakan naive bayes classifier	Metode klasifikasi Naïve Bayes	Dataset Elektabilitas tokoh politik Metode seleksi fitur Chi-Square dan TF-IDF

BAB III
LANDASAN TEORI

3.1 Information Retrieval

Information Retrieval merupakan ilmu yang mempelajari metode dan prosedur untuk menemukan kembali informasi yang tersimpan dari berbagai sumber yang relevan atau koleksi sumber informasi yang dicari atau dibutuhkan. Proses-proses dalam IR dapat berupa pembuatan index (*indexing*), panggilan (*searching*), dan pemanggilan data kembali (*recalling*) (Manning dkk, 2008).

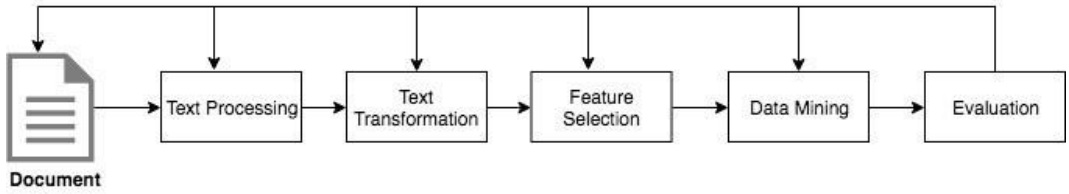
Karena ketersediaan beragam sumber daya di internet, information retrieval sangat subjektif pada jenis media, format data yang didukung oleh media, dan jenis analisis yang diperlukan. Beberapa situs *microblogging* seperti Twitter, Sina-Wiebo menyediakan *Application Programming Interface* (API) mereka untuk mengumpulkan data publik dari situs mereka. Twitter telah menyediakan API REST Twitter untuk mendapatkan data statis seperti informasi profil pengguna, dan Streaming API2 untuk mendapatkan data streaming seperti tweets. Demikian pula dengan Facebook telah menyediakan Facebook Graph API4. API ini membantu dalam mengekstrak postingan dan informasi lain dari Facebook. (Ravi dan Ravi, 2015).

3.2 Text Mining

Text mining adalah suatu bidang didalam data mining dimana penggalian informasi dan pengelolaan sekumpulan dokumen menggunakan tools analisis. Gagasan utama dari text mining adalah mengetahui cakupan atau topik dari permasalahan dalam teks. Pengambilan informasi dari teks (text mining) antara lain dapat meliputi kategorisasi teks atau dokumen, analisis sentimen , pencarian topik yang lebih spesifik, serta *spam filtering*. Text mining penting dalam analisis sentimen sebagai pengidentifikasi emosional pada suatu pernyataan, sehingga banyak studi tentang analisis sentimen dilakukan. (Manning dkk, 2008).

Berbeda dengan data mining yang biasanya memproses data terstruktur, text mining biasanya digunakan untuk memproses *unstructured* atau minimal *semi-structured* data. Akibatnya text mining mempunyai tantangan tambahan yang tidak ditemui di data mining seperti struktur data yang kompleks dan tidak lengkap, arti yang tidak jelas, tidak baku, bahasa yang berbeda serta translasi yang tidak akurat. Oleh karena itu biasanya *Natural Language Processing* (NLP) digunakan untuk memproses unstructured data text tersebut (Adiwijawa, 2006).

Gambar 3.1 menjelaskan bagaima proses yang terjadi di dalam text mining.



Gambar 3.1 Proses dalam Text Mining

Proses *text mining* dimulai dengan *text preprocessing* yaitu pengubahan bentuk dokumen menjadi data terstruktur dengan cara seperti *tokenization*, *stopword removal* , *stemming*, normalisasi dan lain sebagainya. Dari hasil *text processing* selanjutnya dilakukan *text transformation* untuk menemukan fitur fitur yang tersimpan didalam data sesuai kebutuhan yang diperlukan. Setelah diketahui fitur fiturnya selanjutnya dilakukan *feature selection* untuk menentukan fitur yang berpengaruh atau tidak dalam pemodelan data menggunakan perankingan atau pembobotan fitur. Setelah itu, digunakan berbagai algoritma data mining untuk menemukan informasi atau pola yang menarik dari data yang terpilih. Selanjutnya dilakukan evaluasi model apakah pola atau informasi yang dihasilkan bertentangan dengan fakta atau hipotesa sebelumnya (Adiwijawa, 2006)

3.3 Preprocessing

Preprocessing adalah tahap penting yang dilakukan untuk membersihkan data atau merubah data menjadi bentuk data yang terstruktur. Proses membersihkan data meliputi pengecekan data yang tidak konsisten, menghapus

data yang terduplikat dan mengkoreksi kesalahan yang terjadi saat penulisan teks (Wikarsa dan Thair, 2016).

Data mentah yang diperoleh dari berbagai sumber yang sering kali perlu diproses terlebih dahulu sebelum meluncurkan analisis sepenuhnya secara matang. Beberapa langkah preprocessing yang umum adalah: *tokenization*, *stop word removal*, *stemming*, *part of speech* (POS) *tagging*, dan *feature extraction and representation* (Ravi dan Ravi, 2015).

3.3.1 Tokenization

Tokenization adalah proses pemotongan *string input* berdasarkan tiap kata penyusunnya. Pada prinsipnya adalah memisahkan setiap kata yang menyusun suatu dokumen. Pada proses ini dilakukan penghilangan angka, tanda baca dan karakter selain huruf alphabet, karena karakter-karakter tersebut dianggap sebagai pemisah kata (*delimiter*) dan tidak memiliki pengaruh terhadap pemrosesan teks. Pada tahapan ini juga dilakukan proses *case folding*, dimana semua huruf diubah menjadi huruf kecil. *Cleaning* adalah proses membersihkan dokumen dari komponen-komponen yang tidak memiliki hubungan dengan informasi yang ada pada dokumen, seperti *tag html*, *link*, dan *script* (Ling dkk, 2014).

3.3.2 Stemming

Stemming adalah proses pengubahan bentuk kata menjadi kata dasar atau tahap mencari akar kata dari tiap hasil. Dengan dilakukanya proses *stemming* setiap kata berimbuhan akan berubah menjadi kata dasar, dengan demikian dapat lebih mengoptimalkan proses *text mining*. Terdapat 2 poin penting yang dipertimbangkan dalam proses stemming :

1. Kata yang tidak memiliki makna yang sama lebih baik disimpan terpisah
2. Bentuk morfologi dari suatu kata yang memiliki makna dasar yang sama lebih baik dipetakan kedalam stem yang sama

Dua aturan ini cukup baik digunakan dalam teks mining atau *language processing*. *Stemming* biasanya dipertimbangkan sebagai *recall-enhancing device.* Untuk bahasa yang relatif simpel morfologinya, *stemming* tidak dapat bekerja

optimal dibandingkan untuk bahasa yang kompleks morfologinya. Kebanyakan eksperimen *stemming* ini diaplikasikan untuk bahasa inggris (Vijayani dkk, 2015).

3.3.3 Lemmatization

Lemmatization biasanya mengacu pada melakukan sesuatu dengan benar dengan menggunakan kosakata dan analisis morfologi kata-kata, biasanya bertujuan untuk menghapus akhiran huruf saja dan mengembalikan bentuk dasar atau kamus kata, yang dikenal sebagai lemma. *Lemmatization* akan mencoba untuk kembali melihat atau melihat tergantung pada apakah penggunaan token itu sebagai kata kerja atau kata benda. Keduanya mungkin juga berbeda dalam stemming yang paling sering meruntuhkan kata-kata yang berhubungan secara derivatif, sedangkan *lemmatization* umumnya hanya menciutkan bentuk akhiran huruf yang berbeda dari lemma. Pemrosesan *linguistik* untuk *stemming* atau *lemmatization* sering dilakukan oleh komponen tambahan untuk proses pengindeksan, dan sejumlah komponen semacam itu ada, baik komersial maupun *open-source* (Manning dkk, 2008).

3.3.4 Stopword Removal

Stopword removal adalah tahap pemilihan kata kata penting dari hasil token, yaitu kata apa saja yang akan digunakan untuk mewakili dokumen. *Stopword* adalah kata kata yang tidak deskriptif (tidak penting) yang dapat dibuang dengan pendekatan *bag of words* (database kumpulan kata kata yang tidak deskriptif/tidak penting), kemudian kalau hasil tokenizatio itu ada yang merupakan kata tidak penting dalam database tersebut, maka hasil tokenisasi itu dibuang. Biasanya performa *text mining* ataupun *information retrieval* dapat ditingkatkan dengan *stopword removal* ini (Vijayani dkk, 2015).

3.3.5 Lexicon Filter

Leksikon adalah kosakata seseorang, bahasa, atau cabang pengetahuan. Gagasannya adalah menyimpan kumpulan istilah kata (kadang-kadang juga disebut sebagai kosakata atau *leksikon*). Kemudian untuk setiap istilah, kami

memiliki daftar yang mencatat dokumen mana istilah itu terjadi. Setiap item dalam daftar yang mencatat bahwa istilah muncul dalam dokumen secara konvensional disebut sebuah *posting*. Daftar ini kemudian disebut daftar postingan, dan semua daftar postingan yang diambil bersama disebut sebagai postingan. (Manning dkk, 2008)

3.3.6 Regular Expression

Regular expression (regex) adalah sebuah pola penggambaran dari sejumlah text. Nama regex berasal dari salah satu teori matematika dengan nama yang sama. Regex secara jelas memisahkan pola dari teks disekitarnya dan tanda bacanya. (Goyvaerts, 2006). Contoh penggunaan *regular expression* adalah sebagai berikut :

1. “[\(\[].*?[\)\]]” yang merupakan regular expression untuk penghilangan bracket
2. “http\S+” yang merupakan regular expression untuk menghapus URL
3. “[^A-Za-z0-9]+” yang merupakan regular expression untuk membuat dokumen berisi hanya *alpha numeric* yaitu huruf dan angka saja
4. "@\S+" "#\S+" yang merupakan regular expression untuk mengapus mention dan hastag dari dokumen

Regular expression dapat diubah-ubah tergantung setiap kondisi text dalam dokumen yang diinginkan.

3.4 Feature Selection

Feature Selection adalah proses memilih subset dari istilah yang terjadi di set pelatihan dan hanya menggunakan subset ini sebagai fitur dalam klasifikasi teks. Pemilihan fitur melayani dua tujuan utama. Pertama, itu membuat training dan menerapkan *classifier* lebih efisien dengan mengurangi ukuran kosakata yang efektif. Ini sangat penting untuk *classifier* yang, tidak seperti Naive Bayes, berat untuk dilatih. Kedua, *feature selection* sering meningkatkan akurasi klasifikasi dengan menghilangkan fitur *noise*.

3.4.1 Term Frequency-Inverse Document Frequency

Term frequency adalah total frekuensi munculnya sebuah kata term dalam corpus. Untuk menghitung *term frequency*, melibatkan jumlah semua kejadian dari kata dalam semua dokumen dalam corpus. Untuk lebih jelasnya, rumus untuk mencari nilai *term frequency* dapat dilihat pada persamaan (3.1)

$$tf(t_i d_j) = \frac{f_{ij}}{\max(f(w,d) : w \in d)} \quad (3.1)$$

dimana :

f_{ij} = frekuensi kemunculan kata t_i pada dokumen d_j

w = nilai maksimum yang dihitung menggunakan frekuensi dari seluruh term yang muncul pada dokumen d_j

Inverse document frequency (IDF) adalah nilai yang menyatakan bahwa semakin jarang sebuah term muncul dalam dokumen-dokumen yang ada didalam corpus, maka semakin relevan term tersebut. Metode IDF ditambahkan karena term frequency dinilai terlalu sederhana dalam mengukur tingkat pentingnya sebuah term karena tidak melibatkan informasi secara global dalam corpus. IDF dapat membantu dalam mebedakan satu dokumen dengan dokumen-dokumen lainya (Siddiqi dan Sharan, 2015).

$$idf(t_i d_j) = \log\left(\frac{|N|}{1+|\mathrm{d} \in \mathrm{D} : t_i \in \mathrm{d}|}\right) \quad (3.2)$$

dimana :

|N| = jumlah total seluruh dokumen

$|\{d \in D : t_i \in d\}|$ = banyaknya dokumen dimana suatu kata (t_i) muncul

Untuk menghitung TF-IDF, maka hal yang dilakukan adalah mengalikan nilai dari term frequency dengan nilai IDF dari suatu term tersebut. Rumus dari TF-IDF dapat dilihat pada persamaan (3.3)

$$tf - idf(t_i d_j) = tf(t_i d_j) \times idf(t_i d_j) \quad (3.3)$$

dimana :

Tf-idf(t_i, d_j) = bobot TF-IDF kata ke-i dalam dokumen d_j

tf(t_i, d_j) = term frequency kata ke-i dalam dokumen d_j

idf(t_i, d_j) = inverse document frequency kata ke-i dalam dokumen d_j

3.4.2 Chi Square

Seleksi fitur digunakan untuk mereduksi fitur yang tidak relevan dalam proses klasifikasi. Seleksi fitur chi square menggunakan teori statistika untuk menguji independensi sebuah term dengan kategorinya.

Penyeleksian fitur chi square dilakukan dengan cara mengurutkan setiap berdasarkan fitur berdasarkan hasil seleksi fitur chi square dari nilai yang terbesar hingga terkecil. Nilai seleksi fitur chi square yang lebih besar dari nilai signifikan menunjukan penolakan hipotesis independensi. Sedangkan jika dua peristiwa menunjukan dependen, maka fitur tersebut menyerupai atau sama dengan label kategori sesuai pada kategori (Ling dkk, 2014). Rumus chi square dapat dilihat pada persamaan (3.4)

$$X^2(D,t,c) = \sum_{et=\{0,1\}} \sum_{ec=\{0,1\}} \frac{(N_{etec} - E_{etec})^2}{E_{etec}} \qquad (3.4)$$

dimana :

χ^2 (D, t, c) = merupakan nilai Chi Square dari term t untuk kelas c

Netec = observed value (jumlah term t pada kelas c)

Eetec = expected value (jumlah term t pada kelas c)

Sementara contoh untuk mencari nilai salah satu expected value dapat dilihat pada persamaan (3.5)

$$E_{11} = N \times P(t) \times P(c) = N \times \frac{N_{11}+N_{10}}{N} \times \frac{N_{11}+N_{01}}{N} \qquad (3.5)$$

dimana :

N = jumlah dokumen

N_{11} = Jumlah kemunculan term t pada kelas c

N_{10} = Jumlah kemunculan term t pada kelas bukan c

N_{11} = Jumlah kelas c yang memuat term t

N_{01} = Jumlah kelas c yang tidak memuat term t

3.5 Natural Language Processing

Natural Language Processing (NLP) adalah bidang penelitian dan aplikasi yang mengeksplorasi bagaimana komputer dapat digunakan untuk memahami dan memanipulasi teks. Peneliti NLP bertujuan untuk mengumpulkan pengetahuan tentang bagaimana manusia memahami dan menggunakan bahasa sehingga peralatan dan teknik pemasangan dapat dikembangkan untuk membuat sistem komputer memahami dan memanipulasi bahasa alami untuk melakukan tugas yang disukai. Dasar-dasar NLP terletak di sejumlah disiplin ilmu, yaitu. ilmu komputer dan informasi, linguistik, matematika, teknik elektro dan elektronik, kecerdasan buatan dan robotik, dan psikologi. Aplikasi NLP mencakup sejumlah bidang studi, seperti terjemahan mesin, pemrosesan teks dan rangkuman bahasa alami, antarmuka pengguna, multibahasa dan *Cross Language Information Retrieaval* (CLIR), pengenalan suara, kecerdasan buatan dan sistem pakar dan sebagainya. (Jusoh dan Al-Fawareh, 2007)

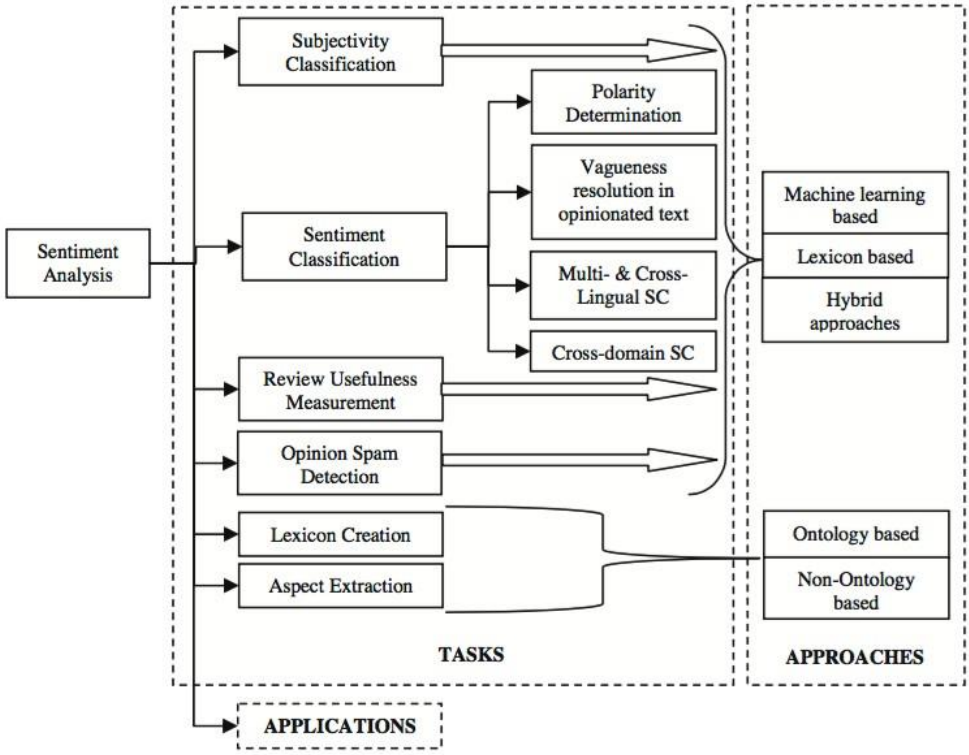
3.6 Support Vector Machines

Support Vector Machines (SVM) secara khusus mendefinisikan kriterianya untuk mencari permukaan keputusan yang jauh dari titik data. Jarak ini dari permukaan keputusan ke titik data terdekat menentukan margin dari classifier. Metode konstruksi ini tentu saja berarti bahwa fungsi keputusan untuk suatu SVM sepenuhnya ditentukan oleh suatu (biasanya kecil) subset data yang mendefinisikan posisi pemisah. Titik-titik ini disebut sebagai *Support Vector* (dalam ruang vektor, titik dapat dianggap sebagai vektor antara titik asal dan titik) (Manning dkk, 2008).

3.7 Sentiment Analysis

Sentiment analysis adalah bidang dalam penelitian yang menggunakan berbagai teknik seperti *Natural Language Processing* (NLP), *Information Retrieval* (IR), dan *Data Mining* (DM) yang digunakan untuk secara sistematis

mengidentifikasi, mengekstrak, mengukur, dan mempelajari keadaan afektif dan informasi yang subjektif. Untuk menghadapi data teks yang tidak terstruktur, metode tradisional NLP sepert *information retrieval* dan *information extraction* muncul. Untuk mendapatkan pengertian dari teks yang diekstraksi, banyak upaya penelitian telah dilakukan dalam beberapa tahun terakhir yang mengarah ke *Automated SA*, area penelitian NLP yang lain (Ravi dan Ravi, 2015).



Gambar 3.2 Proses Sentiment Analysis

Mengacu pada gambar 3.2, sentiment analysist bukanlah masalah tunggal sebaliknya itu adalah masalah yang beraneka ragam. Berbagai langkah diperlukan untuk melakukan *opinion mining* dari teks yang diberikan, karena teks untuk penambangan opini berasal dari beberapa sumber daya dalam beragam jenis. Akuisisi data dan preprocessing data adalah sub-bagian yang paling umum yang diperlukan untuk penambangan teks dan SA.

3.8 Pengujian

Pengujian dilakukan untuk menghitung performa dari sistem atau dalam penelitian ini performa klasifikasi sentimen. Untuk menghitung performa klasifikasi, data latih digunakan untuk melakukan validasi. Validasi dapat dilakukan dengan menggunakan metode yang dinamakan *k-fold cross validation*.

3.8.1 K-Fold Cross Validation

Cross Validation merupakan salah satu metode pengujian dalam *data mining* dan *machine learning*. Dalam melakukan *cross validation*, dataset dibagi menjadi beberapa bagian atau dalam hal ini disebut *fold* secara acak. Satu bagian dari hasil pembagian tersebut digunakan sebagai data pengujian dan sisanya digunakan sebagai data pelatihan. Pengujian dilakukan sejumlah bagian yang ada dengan bergantian menggunakan bagian yang berbeda sebagai data pengujian. Hasil pengujian yang dihitung adalah hasil keseluruhan yaitu merata-rata hasil dari keseluruhan pengujian yang dilakukan (Witten dkk, 2011).

Cross validation yang dilakukan dengan pembagian sebanyak *k* disebut sebagai *k-fold cross validation*. Jumlah bagian yang umum digunakan untuk pengujian adalah sebanyak sepuluh. Pengujian banyak dilakukan dengan menggunakan *10-fold cross validation* dikarenakan dari hasil percobaan menunjukkan *10-fold cross validation* memberikan perkiraan kesalahan dari klasifikasi yang terbaik. Pengujian *k-fold cross validation* yang dilakukan sekali sering tidak memberikan perkiraan kesalahan yang bisa diandalkan. Hal ini disebabkan pengujian yang dilakukan beberapa kali dengan parameter yang sama dapat memberikan hasil yang berbeda yang disebabkan oleh unsur *random* yang ada pada saat pembagian data. Oleh karena itu untuk mendapatkan hasil yang dapat diandalkan pengujian harus dilakukan berkali-kali dengan hasilnya merupakan rata-rata dari semua hasil pengujian.

Contoh penggunaan *k-fold cross validation* dengan k=5 bisa dilihat pada gambar 3.3 berikut

	Model 1	Model 2	Model 3	Model 4	Model 5
Fold 1	Data Test	Data Train	Data Train	Data Train	Data Train
Fold 2	Data Train	Data Test	Data Train	Data Train	Data Train
Fold 3	Data Train	Data Train	Data Test	Data Train	Data Train
Fold 4	Data Train	Data Train	Data Train	Data Test	Data Train
Fold 5	Data Train	Data Train	Data Train	Data Train	Data Test

Gambar 3.3 : Ilustrasi k-fold cross validation dengan k=5

3.9 Perhitungan Performa

Model klasifikasi adalah pemetaan dari suatu input data menjadi suatu output yang merupakan prediksi kelas. Klasifikasi yang hanya menghasilkan dua kelas sebagai outputnya disebut klasifikasi biner. Kedua kelas tersebut sering kali merupakan kelas positif dan kelas negatif. Terdapat empat kemungkinan yang terjadi dari proses pengklasifikasian biner seperti yang dijabarkan oleh Fawcett (2006):

- True Positive (TP) : Prediksi menghasilkan true, dan kejadiannya true. Sebagai contoh jika model memprediksi seseorang mendapat sentimen positif dan pada kenyataannya mereka memiliki sentimen positif.
- False Positive (FP) : Prediksi menghasilkan true, dan kejadiannya false. Sebagai contoh jika model memprediksi seseorang mendapat sentimen positif dan pada kenyataannya mereka memiliki sentimen negatif.
- True Negative (TN) : Prediksi menghasilkan false, dan kejadiannya false. Sebagai contoh jika model memprediksi seseorang mendapat sentimen negatif dan pada kenyataannya mereka memiliki sentimen negatif.
- False Negative (FN) : Prediksi menghasilkan false, dan kejadiannya false. Sebagai contoh jika model memprediksi seseorang mendapat sentimen negatif dan pada kenyataannya mereka memiliki sentimen positif.

Hasil klasifikasi biner pada suatu data set dapat direpresentasikan pada suatu matriks 2x2 yang disebut confusion matrix seperti yang ditunjukkan pada Gambar 3.4

Gambar 3.4: Ilustrasi Confussion Matrix

Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi yang dapat dijabarkan sebagai berikut menurut Fawcett (2006):

Accuracy (Akurasi) : perbandingan kasus yang diidentifikasi benar dengan jumlah semua kasus. Persamaan 3.6 menunjukkan rumus untuk accuracy.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.6)$$

Precision/PPV : proporsi kasus yang diidentifikasi dengan benar sebagai milik kelas ‘a’ di antara semua kasus yang diklasifikasikan oleh pengklasifikasi bahwa mereka termasuk dalam kelas ‘a’. Persamaan 3.7 menunjukkan rumus untuk menghitung precision.

$$Precison = \frac{TP}{TP+FP} \quad (3.7)$$

Recall/Sensitivity : proporsi kasus yang diidentifikasi dengan benar sebagai milik kelas ‘a’ di antara semua kasus yang benar-benar termasuk dalam kelas ‘a’. Persamaan 3.8 menunjukkan rumus untuk menghitung recall.

$$Recall = \frac{TP}{TP+FN} \quad (3.8)$$

F1 score/F-measure : bobot rata-rata antara precision dan recall. Persamaan 3.9 menunjukkan rumus F1 *score*.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.9)$$

Nilai TP, TN, FP, dan FN yang ada pada ketiga rumus diatas masing-masing merupakan nilai *true positive, true negative, false positive*, dan *false negative*. Keempat nilai akurasi, precision, recall, dan F1 score ditampilkan dalam bentuk persentase.

3.10 Positive Versus Total

Positive versus Total (PvT) adalah cara untuk menghitung elektabilitas tokoh dengan membandingkan jumlah sentiment positif dari masing-masing tokoh dengan jumlah total sentimen (positif dan negatif) (Ramteke dkk, 2016). Rumus PvT dapat dilihat pada persamaan (3.11).

$$PvT(x) = \frac{Positive(x)}{Total(x)} \quad (3.10)$$

dimana :

Positive(x) = Jumlah nilai sentimen positif pada tokoh *x*

Total(x) = Jumlah nilai keseleruhan positif dan negatif pada tokoh *x.*

3.11 Share Of Volume

Share of Volume (SoV) adalah cara menghitung elektabilitas tokoh politik dengan membandingkan jumlah sentimen positif seorang tokoh politik dengan total sentimen positif keseluruhan tokoh politik. SoV memiliki keuntungan bahwa hasilnya bisa dibandingkan dengan mudah dengan hasil presentasi polling (Bermingham dan Smeaton, 2011). Rumus perhitungan SoV ditunjukan pada persamaan (3.11)

$$SoV(x) = \frac{Positive(x)}{\sum_{i=1}^{n} Positive(i)} \quad (3.11)$$

dimana :

Positive(x) = Jumlah nilai sentimen positif pada tokoh *x*

Positive(i) = Jumlah semua nilai positif pada semua tokoh.

BAB IV
METODOLOGI PENELITIAN

4.1 Alat dan Bahan

Alat dan bahan yang digunakan dalam penelitian ini yaitu:

1. Laptop Macbook Pro Intel Core 2 Duo RAM 4 GB
2. Sistem Operasi MacOS El Capitan 10.11.6
3. Jupyter Notebook 5.7.0
4. Python 3.6.3
5. Google Colab

4.2 Tahapan Penelitian

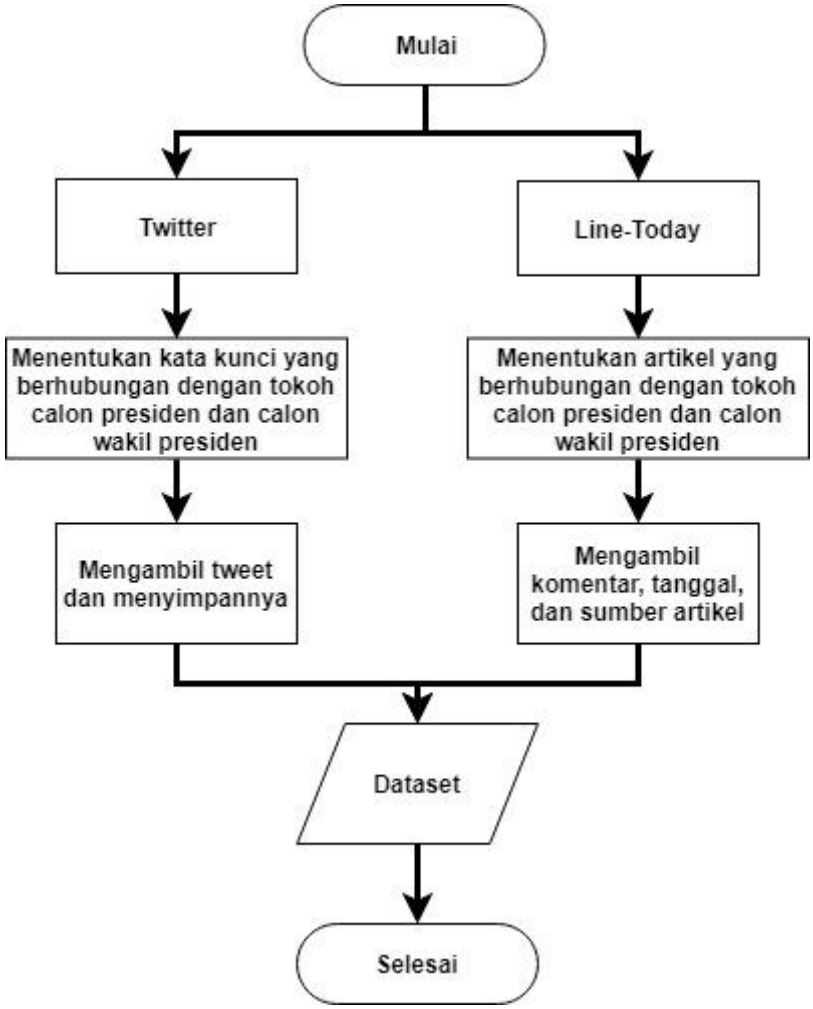
Tahapan yang dilakukan dalam penelitian ini secara umum adalah :

4.2.1 Studi Literatur

Pada tahap ini dilakukan pengumpulan informasi informasi yang mendukung penelitian ini. Informasi yang telah dikumpulkan mencakup penjelasan mengenai sentiment analysis, scrapping data, multinomial naïve bayes, support vector machine, text mining, metode evaluasi k-fold cross validation dan confusion matrix untuk menguji akurasi model. Informasi tersebut didapatkan dari berbagai sumber seperti jurnal ilmiah, paper, website, buku, skripsi dan sebagainya.

4.2.2 Pengumpulan Data

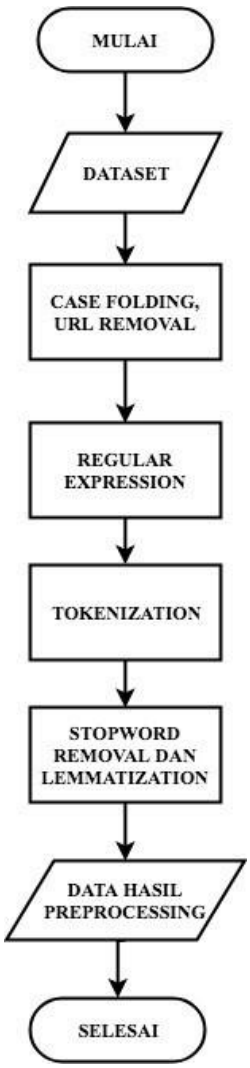
Penelitian ini menggunakan data dari Twitter dan Line-Today yang diambil menggunakan beberapa tools untuk scrapping. Data set akan dibagi menjadi data training dan data testing dengan jumlah yang berbeda. Setiap data latih kemudian dilabeli secara manual untuk sentimen positif atau negatifnya. Rangkaian dari proses pengumpulan data ditunjukan pada gambar 4.1.



Gambar 4.1 : Diagram Alur Scrapping Data

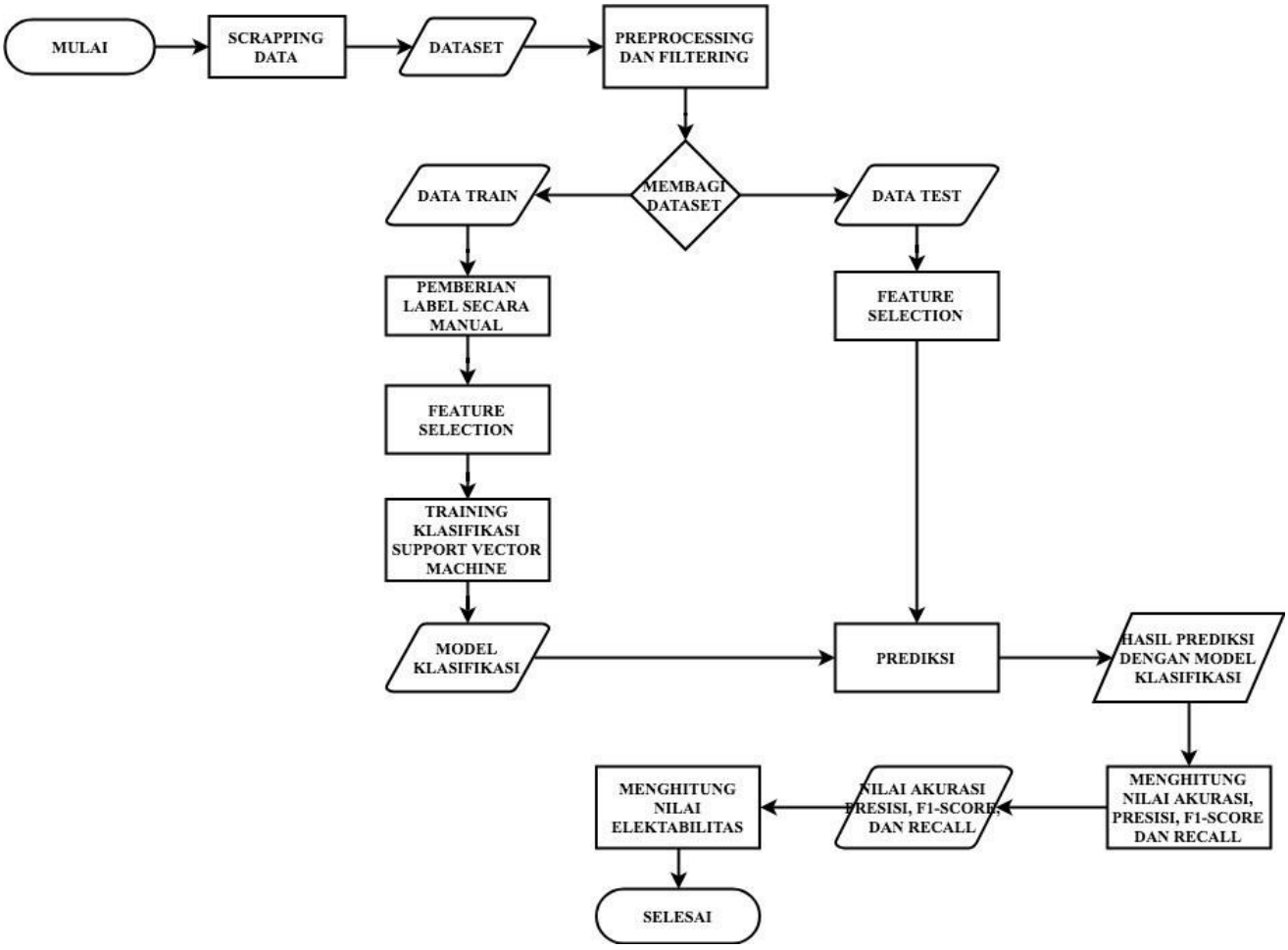
4.2.3 Perancangan Sistem

Proses preprocessing mengubah data yang tidak terstruktur menjadi data terstruktur agar lebih mudah diolah. Hal ini dilakukan untuk agar data yang memiliki banyak noise dapat diproses dan dianalis oleh sistem. Proses yang dilakukan antara lain penghapusan URL, case folding untuk merubah menjadi huruf kecil, stopword removal, lexicon filtering dan stemming untuk mendapatkan kata dasar dan dilanjutkan dengan pembobotan. Rangkaian dari proses preprocessing ditunjukan pada gambar 4.2.



Gambar 4.2 Diagram Alur Preprocessing

Proses klasifikasi akan menggunakan model support vector machine (pada gambar 4.3). Hal ini nantinya akan diuji nilai akurasi, presisi, recall, F-1. Lalu akan diuji kembali dengan Hyperparameter Tuning pada model classifier pada kedua model untuk meningkatkan nilai pengujian.



Gambar 4.3 : Diagram alur klasifikasi dengan Support Vector Machine

Setelah semua tweet dan berita terklasifikasi positif dan negatif, hasil klasifikasi sentimen tersebut digunakan untuk menghitung nilai elektabilitas dari masing-masing tokoh politik. Perhitungan nilai elektabilitas tokoh politik dilakukan dengan dua cara yaitu :

1. *Positive versus total*, dengan membagi jumlah sentimen positif dan jumlah total sentimen (jumlah data) dari tokoh politik tersebut.
2. *Share of Volume*, dengan membagi jumlah sentimen positif seorang tokoh politik dengan total sentimen positif dari keseluruhan tokoh politik.

Perhitungan nilai elektabilitas dilakukan untuk data yang menggunakan seleksi fitur TF-IDF dan Chi Square dan tanpa seleksi fitur. Hasil perhitungan nilai

elektabilitas PvT dan SOV akan dirata-rata untuk mendapatkan nilai elektabilitas akhir dari masing-masing tokoh politik.

4.2.4 Implementasi Sistem

Pada tahap ini dilakukan implementasi algoritma dan skenario yang sudah dirancang dalam bentuk program computer. Program tersebut akan dibuat dengan menggunakan bahasa pemrograman Python3 dengan menggunakan Jupyter Notebook atau Google Colab.

4.2.5 Pengujian

Pengujian akan dilakukan dengan menggunakan metode 10-fold *cross validation*. Secara *stratified* data latih akan dibagi menjadi 10 bagian. Kemudian satu bagian akan diujikan sebagai data uji, sedangkan bagian lainya akan digunakan sebagai data latih. Pengujian dilakukan sebanyak sepuluh kali dengan digunakan bagian yang berbeda-beda sebagai data ujinya. Hasil tersebut kemudian di rata rata dari setiap pengujian yang dilakukan. Hasil pengujian tersebut kemudian dimasukan kedalam *confussion matrix* dan dihitung akurasinya, presisi, *recall*, dan f-1 *score* nya. Pada akhir klasifikasi akan didapatkan label sentimen pada keseluruhan data untuk digunakan mengukur elektabilitas tokoh politik.

4.2.6 Penulisan Laporan

Tahapan penulisan laporan dilakukan sejalan dengan berjalannya penelitian.

BAB V
JADWAL PENELITIAN

Jadwal penelitian dibuat sebagai acuan dalam menyelesaikan penelitian sesaui dengan tahapan dan target yang harus dicapai dalam waktu yang telah direncanakan. Jadwal penelitian tertera pada Tabel 5.1.

Tabel 5.1 Rencana Jadwal Penelitian

Kegiatan	2019																			
	Januari				Februari				Maret				April				Mei			
Minggu Ke-	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Studi Literatur																				
Pengumpulan Data																				
Perancangan Sistem																				
Implementasi Sistem																				
Pengujian Sistem																				
Penulisan Laporan																				

DAFTAR PUSTAKA

Abramowitz, A. I. (1989). Viability, electability, and candidate choice in a presidential primary election: A test of competing models. The Journal of Politics, 51(4), 977-992.

Adiwijawa, I. (2006). Text Mining dan Knowledge Discovery. EMC Coporation.

Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L., & Desouza, K. C. (2016). An analysis of sentiments on facebook during the 2016 US presidential election. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp.795-802). IEEE.

Bermingham, A., & Smeaton, A. (2011). On using Twitter to monitor political sentiment and predict election results. In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011) (pp. 2-10).

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

Goyvaerts, J. (2006). Regular Expressions: The Complete Tutorial. Lulu Press. Hakimi, F. D. D. (2018). Sistem analisis sentimen publik tentang opini pemilihan Kepala Daerah Jawa Timur 2018 pada dokumen twitter menggunakan naive bayes classifier.

Hidayatullah, A. F., & Azhari, A. S. (2015). Analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada Twitter. In Seminar Nasional Informatika (SEMNASIF) (Vol. 1, No. 1).

Jusoh, S., & Al-Fawareh, H. M. (2007). Natural language interface for online sales systems. In Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on (pp. 224-228). IEEE.

Lestari, A. R. T., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji. Jurnal

Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

Ling, J., Kencana, I. P. E. N., & Oka, T. B. (2014). Analisis Sentimen Menggunakan Metode Naïve Bayes Classifier Dengan Seleksi Fitur Chi Square. E-Jurnal Matematika, 3(3), 92-99.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). Introduction to information retrieval (Vol. 39). Cambridge University Press.

Prasetyo, N. D. (2014). Tweet-based election prediction (Doctoral dissertation, Ph. D. dissertation, TU Delft, Delft University of Technology).

Ramteke, J., Shah, S., Godhia, D., & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. In Inventive Computation Technologies (ICICT), International Conference on (Vol. 1, pp. 1-5). IEEE.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.

Razzaq, M. A., Qamar, A. M., & Bilal, H. S. M. (2014). Prediction and analysis of Pakistan election 2013 based on sentiment analysis. In Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 700-703). IEEE Press.

Saputra, N., Adji, T. B., & Permanasari, A. E. (2015). Analisis sentimen data presiden Jokowi dengan preprocessing normalisasi dan stemming menggunakan metode naive bayes dan SVM. Jurnal Dinamika Informatika, 5(1).

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications, 109(2).

Sukendar, M. U., Sos, S., & Kom, M. I. (2017). Pemilihan Presiden, Media Sosial Dan Pendidikan Politik Bagi Pemilih Pemula. Jurnal IKON Prodi D3 Komunikasi Massa–Politeknik Indonusa Surakarta Vol, 1(5).

Suryotomo, R. (2018). Analisis Sentimen Untuk Mengetahui Elektabilitas Tokoh Politik Menggunakan Metode Multinomial Naive Bayes.

Virgo, F. G. (2018). Analisis Sentimen Dan Deteksi Buzzer Dalam Prediksi Pilkada Dki Jakarta 2017.

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 90-94). Association for Computational Linguistics.

Wicaksono, A. J. 2016. A proposed method for predicting US presidential election by analyzing sentiment in social media. In Science in Information Technology (ICSITech), 2016 2nd International Conference on (pp.276-280). IEEE.

Wikarsa, L., & Thahir, S. N. (2015, November). A text mining application of emotion classifications of Twitter's users using Naïve Bayes method. In Wireless and Telematics (ICWT), 2015 1st International Conference on (pp. 1-6). IEEE.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. The Annals of Applied Statistics, 5(4), 2493-2518.

Yuniarni, S. (2018). Indonesia Had 143m Internet Users in 2017: APJII. Retrieved November 19, 2018, from https://jakartaglobe.id/business/indonesia-143m-internet-users-2017-apjii/