BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

Analisis sentimen adalah sebuah teknik untuk mengetahui opini masyarakat terhadap suatu subyek tertentu yang didapat dari sebuah kumpulan data. Dengan pertumbuhan teknologi informasi yang semakin canggih berpengaruh pada perubahan cara manusia dalam berkomunikasi terhadap sesamanya. Penggunaan media sosial banyak digunakan oleh masyarakat umum untuk berkomunikasi atau berekspresi menyampaikan opininya. Penelitian tentang analisis sentimen sendiri sampai sekarang sudah banyak dilakukan.

Pada jurnal yang ditulis oleh Liu menjelaskan bahwa analisis sentimen adalah bidang studi yang menganalisis pendapat, sentimen, evaluasi, sikap, dan emosi seseorang dari bahasa tulisan. Analisis sentimen merupakan salah satu bidang penelitian yang paling aktif dalam pemrosesan bahasa alami dan juga banyak dipelajari dalam *data mining*, *web mining*, dan *text mining* (Liu, 2012). Selain di bidang ilmu komputer, analisis sentimen juga berguna di bidang lainnya, seperti ilmu manajemen dan ilmu sosial. Menurut Prabowo dan Thelwall sentimen adalah suatu kata atau kalimat yang ditemukan dalam komentar, umpan balik atau kritik memberikan indikator berguna untuk berbagai tujuan. Sentimen ini dapat dikategorikan menjadi dua kategori: positif dan negatif, atau ke skala

n-point, misalnya, sangat bagus, bagus, memuaskan, buruk, sangat buruk. Dalam hal ini, analisis sentimen dapat diartikan sebagai klasifikasi di mana masing-masing kategori mewakili sentimen (Prabowo & Thelwall, 2009).

Analisis sentimen juga menggunakan algoritma untuk mengolah dan melakukan klasifikasi terhadap data yang dibangun. Terdapat banyak algoritma yang dapat digunakan dalam penelitian analisis sentimen. Ada sepuluh algoritma terbaik yang biasa digunakan, di antaranya adalah *C4.5*, *The K-Means*, *Support Vector Machine*, *Apriori*, *Maximum Entropy PageRank*, *AdaBoost*, *k-nearest neighbor*, *Naive Bayes*, *CART* (Wu et al., 2008). Pada penelitian ini penulis menggunakan algoritma *Support Vector Machine*, algoritma tersebut dipilih karena dapat melakukan klasifikasi data yang lebih dari dua kelas dengan konsep multiSVM dan dapat meminimalisir *error* dalam *training set.* Dalam penelitiannya yang dilakukan oleh Nugroho, Witarto, dan Handoko, mereka menggunakan *Support Vector Machine* karena penelitian mereka menggunakan sampel data yang sedikit dan pada *Support Vector Machine* memiliki kelebihan dalam menangani *Curse of dimensionality* yang didefinisikan sebagai masalah yang dihadapi suatu metode *pattern recognition* dalam mengestimasikan parameter karena jumlah sampel yang sedikit (Nugroho, Witarto, & Handoko, 2003).

Support Vector Machine juga digunakan oleh Zainuddin dan Selamat dalam penelitiannya. Mereka

menjelaskan mengenai penggunaan metode *Support Vector Machine* pada dataset *benchmark* untuk melatih sentimen *classifier*. Dalam penelitian tersebut menggunakan banyak fitur dan *Support Vector Machine* menghasilkan rata-rata AUC diatas 0,8. Dari berbagai fitur yang diuji, *Support Vector Machine* menghasilkan akurasi yang tinggi jika fitur yang digunakan sedikit (Zainuddin & Selamat, 2014).

Untuk mengukur performa dari algoritma yang dipakai dilakukan penelitian menggunakan dua algoritma. Penggunaan dua algoritma pernah dilakukan dalam penelitian mengenai kebiasaan masyarakat dalam menilai tokoh publik dengan menggunakan data dari Twitter. Dalam penelitian ini sebanyak 1329 data *tweet* dilabeli secara manual, kemudian klasifikasi menggunakan algoritma *Naïve Bayes Classifier* dan *Support Vector Machine*. Mereka melakukan pengujian menggunakan *RapidMiner* dengan fitur *term frequency* dan fitur TF-IDF. Hasil Pengujian menggunakan fitur term frequency diperoleh akurasi sebesar 79,91%, dan menggunakan fitur TF-IDF diperoleh akurasi sebesar 79,68%. Hasil pengujian klasifikasi *tweet* menggunakan *Naïve Bayes Classifier dengan fitur term frequency* diperoleh akurasi sebesar 73,81% sedangkan dengan fitur TF-IDF diperoleh akurasi sebesar 71,11% dan hasil pengujian menggunakan *Support Vector Machine* dengan fitur *term frequency* diperoleh akurasi sebesar 83,14% sedangkan dengan fitur TF-IDF diperoleh akurasi sebesar 82,69%. Kesimpulan dari

penelitian ini fitur *term frequency* lebih baik dari TF-IDF dan metode *Support Vector Machine* memiliki akurasi lebih baik dari *Naïve Bayes Classifier* (Hidayatullah & Sn, 2014).

Dalam penelitian ini penulis juga menggunakan media Twitter sebagai sumber *dataset*. Penggunaan Twitter juga popular dalam penelitian analisis sentimen sebelumnya. Penggunaan Twitter sebagai *dataset* juga dikarenakan Twitter berisi sejumlah besar pesan singkat yang sangat banyak yang diciptakan oleh pengguna *microblogging* ini. Isi pesan bervariasi dari pemikiran pribadi hingga pernyataan publik (Pak & Paroubek, 2016). Pada jurnal yang ditulis Novantirani, Sabariah, dan Effendy mereka meneliti mengenai sentimen masyarakat terhadap transportasi darat dalam kota menggunakan data Twitter. Dalam penelitian ini dataset yang dibangun berupa contoh angkutan kendaraan darat sebanyak empat kendaraan, yaitu: Angkot, Kopaja, Metromini, dan Transjakarta. Metode yang digunakan untuk klasifikasi dalam penelitian ini adalah *Support Vector Machine*. Hasil dari pengujian penelitian ini memperoleh rata-rata akurasi di kisaran 67,83% dan hasil tertinggi adalah 78,125% pada dataset Transjakarta (Novantirani, Sabariah, & Effendy, 2015). Pada penelitian ini, mereka juga menganalis faktor opini terhadap masing-masing angkutan kendaraan darat untuk menemukan opini yang paling tinggi frekuensinya terhadap angkutan kendaraan darat tersebut. Analisa faktor opini juga akan dilakukan

pada penelitian yang penulis lakukan untuk menemukan opini apa yang paling tinggi frekuensinya terhadap objek yang dianalisa.

Dalam penelitian ini penulis menganalisa opini terhadap hasil dari Pilkada DKI Jakarta 2017. Pada penelitian sebelumnya pernah dilakukan analisa sentimen terhadap calon gubernur DKI Jakarta 2017 oleh Marpaung. Dalam penelitian tersebut data yang digunakan adalah data dari media sosial Twitter. Data-data yang dikumpulkan berkaitan dengan pemilihan Gubernur DKI Jakarta 2017, kemudian klasifikasi menggunakan metode *Naïve Bayes Classifier* untuk mengetahui sentimen positif, negatif, atau netral. Hasil akhir dari penelitian ini adalah klasifikasi sentimen dominan pada sentimen positif terhadap Basuki Tjahaja Purnama atau Ahok (Marpaung, 2017).

Analisa terhadap calon gubernur DKI Jakarta 2017 juga dilakukan Buntoro pada penelitian ini juga menggunakan data Twitter. Dataset yang digunakan ada tiga, yaitu data Ahok, data Anies, dan data AHY. Algoritma yang digunakan adalah *Support Vector Machine* dan *Naïve Bayes.* Hasil akhir dari penelitian ini adalah akurasi *Support Vector Machine* lebih tinggi pada data Anies tapi Akurasi *Naïve Bayes* lebih tinggi pada data AHY. Hal ini dikarenakan jumlah pembagian sentimen yang tidak seimbang (Buntoro, 2016). Pada penelitian Analisa terhadap opini Pilkada DKI Jakarta 2017 Pada Dokumen Twitter Berbahasa Indonesia menggunakan algoritma *Naïve Bayes*. Dari hasil

pengujian akurasi, diperoleh 68,52% untuk kondisi pembobotan tekstual, 75,93% untuk pembobotan non-tesktual, dan 74,81% untuk kondisi penggabungan dengan nilai konstanta 0,5 untuk tekstual dan 0,5 untuk non-tekstual (Rossi, Lestari, Perdana, & Fauzi, 2017).

Dalam penelitian ini, penulis membangun *dataset* sentimen, dalam pembentukan *dataset* penulis menerapkan pemrosesan bahasa natural meliputi penhapusan *stopwords*, *stemming*, dan pengelompokan kelas atribut menggunakan bantuan kamus *SentiStrength* sebagai acuan untuk melabelkan kata yang berkonotasi positif, negatif, dan netral. Setelah *dataset* terbentuk, penulis menerapkan algoritma *Support Vector Machine.* Hasil dari algoritma *Support Vector Machine* dibandingkan dengan metode lainnya.

Tabel 2.1 Algoritma Support Vector Machine untuk klasifikasi data

Item Pembanding	Nugroho, Witarto, & Handoko, 2003	Zainuddin & Selamat, 2014	Novantirani, Sabariah, & Effendy, 2015
Judul Penelitian	*Support Vector Machine* - Teori dan Aplikasinya dalam Bioinformatika1	Sentiment Analysis Using *Support Vector Machine*	Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode *Support Vector Machine*
Tujuan	membahas teori dasar *Support Vector Machine* dan aplikasinya dalam bioinformatika, khususnya pada analisa ekspresi gen yang diperoleh dari analisa microarray.	Eksperimen terhadap seleksi fitur Chi-Square untuk memberikan peningkatan yang signifikan pada akurasi klasifikasi	Mengetahui opini masyarakat terhadap angkutan kendaraan darat
Metode	*Support Vector Machine*	*Support Vector Machine*	*Support Vector Machine*
Hasil	*Support Vector Machine* mampu mengklasifikan suatu pasien terkena penyakit kanker atau tidak, berdasarkan hasil analisa microarray terhadap sel pasien tersebut	*Support Vector Machine* mampu menghasilkan akurasi yang tinggi dengan AUC rata-rata diatas 0,8 dan pemilihan fitur chi-square secara signifikan meningkatkan akurasi klasifikasi	*Support Vector Machine* mampu menghasilkan akurasi dengan kisaran 67,83% dan hasil tertinggi adalah 78,125%
Sasaran	Bidang Medis	Online Movie	Angkutan Kendaraan Darat

Tabel 2.2 Perbandingan algoritma Support Vector Machine dengan metode lainnya

Item Pembanding	Hidayatullah & Sn, 2014	Buntoro, 2016
Judul Penelitian	Analisis Sentimen Dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter	Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter
Tujuan	menganalisis tweet berbahasa Indonesia yang membicarakan tentang tokoh publik	melakukan riset atas opini masyarakat yang mengandung sentimen positif, netral atau negati
Metode	*Naïve Bayes Classifier* dan *Support Vector Machine*	*Naïve Bayes Classifier* dan *Support Vector Machine*
Hasil	*Support Vector Machine* menghasilkan akurasi performansi yang lebih baik daripada metode *Naive Bayes*	*Naïve Bayes Classifier* lebih tinggi akurasinya untuk klasifikasi sentimen Tweet Bahasa Indonesia dibandingkan dengan metode klasifikasi *Support Vector Machine*
Sasaran	Tokoh Publik	Calon Gubernur DKI Jakarta 2017

2.2. Landasan Teori

a. Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Istilah data mining memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. *Data mining*, sering juga disebut sebagai *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Ridwan et al., 2013). Dalam *Data Mining* ada tiga karakteristik data (Leslie, Spits, Lumban, Trisetyarso, & Abdurachman, 2017), yaitu:

1. *Supervised*, adalah variabel atau data yang berlabel.
2. *Semi Supervised*, adalah variabel atau data yang beberapa berlabel dan beberapa tidak berlabel.
3. *Unsupervised*, adalah variabel yang tidak berlabel.

b. Analisis Sentimen

Analisis sentimen sendiri atau juga biasa disebut dengan *opinion mining* adalah salah satu

bagian dari *text mining*. Bidang ini melakukan studi mengenai opini orang-orang, sentimen, evaluasi, tingkah laku dan emosi terhadap suatu entitas seperti produk, layanan, organisasi, individu, permasalahan, topik, acara dan atribut-atributnya. Analisis sentimen sangatlah berguna untuk menganalisis komentar-komentar di Twitter tadi untuk kemudian diterjemahkan menjadi sesuatu yang lebih bermakna, salah satunya dalam bentuk *rating*. Dalam dunia bisnis *rating* menjadi sangat penting karena merupakan salah satu indikator kesuksesan. Di sisi lain, rating masih mejadi komoditas monopoli beberapa perusahaan seperti Nielsen, sehingga objektivitasnya menjadi kurang. Celah ini lah yang kemudian dimanfaatkan penulis untuk mencoba mengaplikasikan analisis sentimen pada Twitter untuk membuat sistem rating berdasar komentar. (Monarizqa, Nugroho, & Hantono, 2014).

c. Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas yang tidak diketahui dari suatu objek. Dalam klasifikasi terdapat dua proses, yaitu proses *training* dan proses *testing* (Bertalya, 2009). Pada proses training menggunakan *training set* yang telah diketahui label-labelnya untuk membangun model. Kemudian

proses *testing* untuk menguji keakuratan model yang telah dibangun saat proses *training*.

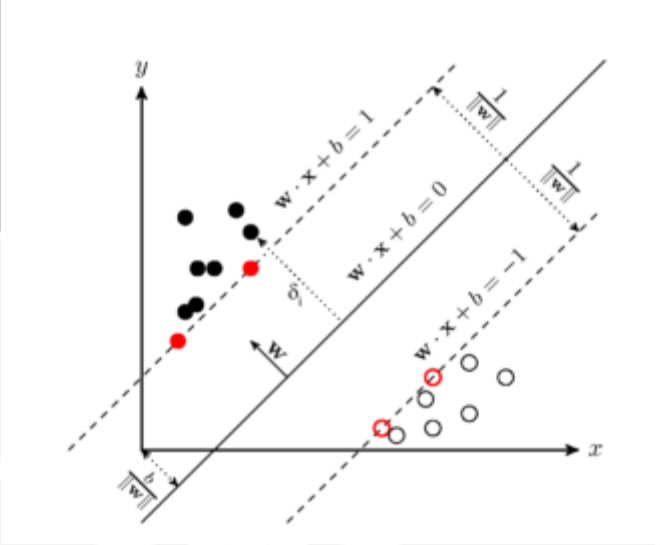
d. Twitter

Twitter merupakan sebuah media sosial yang memberikan layanan dari *microblogging* untuk memberikan fasilitas terhadap pengguna dalam mengirim dan membaca pesan dalam berupa *tweets*. *Microblogging* merupakan sebuah layanan berbasis web dimana penggunanya dapat menulis status dalam berupa teks, mengunggah gambar atau video secara *online* dan *real time*. *Tweet* merupakan sebuah teks tulisan yang memiliki batasan mencapai 140 karakter. Pengguna yang menuliskan status kedalam *tweets* dapat dilihat secara publik, namun juga dapat mengirim pesan melalui daftar *follower* mereka saja *(direct message)*. *Follower* merupakan pengguna lain yang dikenal oleh pengguna yang dapat disebut dengan pengikut.

e. *Support Vector Machine*

Support Vector Machine merupakan salah satu metode klasifikasi dengan menggunakan *machine learning* (supervised learning) yang memprediksi kelas berdasarkan model atau pola dari hasil proses training. Klasifikasi dilakukan dengan mencari hyperplane atau garis pembatas (decision boundary) yang memisahkan antara suatu kelas dengan kelas lain (Novantirani et al., 2015), yang dalam kasus ini garis tersebut berperan memisahkan tweet

bersentimen positif (berlabel +1) dengan *tweet* bersentimen negatif (berlabel -1). *Support Vector Machine* melakukan pencarian nilai *hyperplane* dengan menggunakan *support vector* dan nilai margin.



Gambar 2.1 SVM berusaha untuk menemukan hyperplane terbaik yang memisahkan kedua kelas -1 dan +1 (Susilowati et al., 2015)

Gambar 2.1 memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah *class*: +1 dan −1. *Pattern* yang tergabung pada *class* −1 disimbolkan dengan warna putih sedangkan *pattern* pada *class* +1, disimbolkan dengan warna hitam. Kedua kelas tersebut dipisahkan oleh garis yang disebut *hyperplane* Persamaan garis *hyperplane* adalah:

$$w.x + b = 0 \quad (2.1)$$

(w adalah normal bidang dan b adalah bias)

Pattern yang memiliki jarak paling dekat dengan hyperplane disebut *support vector*, disimbolkan dengan warna merah. *Support Vector Machine* memisahkan data menggunakan *hyperplane* dengan

batas antar kelas terbesar. Maka dari itu dibentuk garis pembatas yang sejajar dengan *support vector* semua kelas. Persamaan yang terbentuk dari garis pembatas tersebut adalah

Garis pembatas positif = w.x + b = 1 (2.2)

Garis pembatas negatif = w.x + b = -1 (2.3)

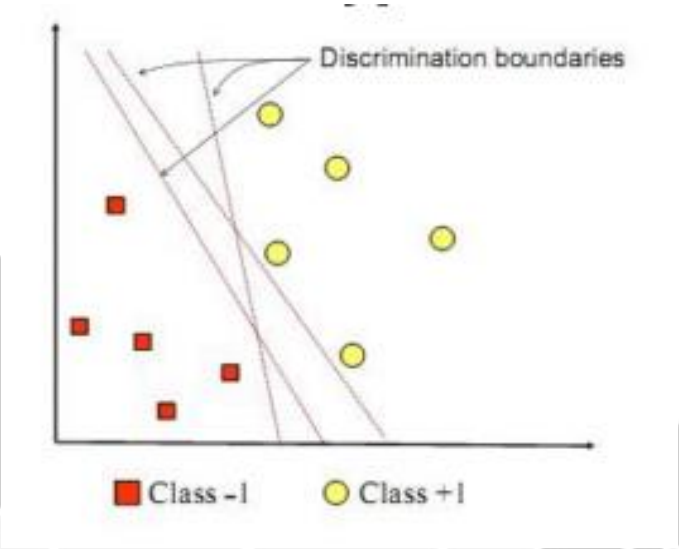
Dari persamaan tersebut dapat diketahui data yang digolongkan positif adalah data yang memiliki nilai persamaan (2.2) sedangkan data yang digolongkan negatif adalah data yang memiliki nilai persamaan (2.3). Dari hasil tersebut dapat dibentuk pertidaksamaan diantara kedua gari pembatas dengan rumus:

$$yi\ (xi.w + b) - 1 \geq 0 \quad (2.4)$$

Nilai batas pada setiap bidang pembatas ke *hyperplane* adalah $\frac{1}{||w||}$, maka dapat ditentukan nilai batas gabungan adalah $\frac{1}{||w||}, + \frac{1}{||w||}, = \frac{2}{||w||},$. Selanjutnya nilai batas ini akan dimaksimalkan, memaksimalkan nilai batas berarti meminimalkan nilai dari ||w|| sebagai penyebut. Jika kedua bidang pembatas sesuai dengan pertidaksamaan diatas, maka pencarian *hyperplane* dengan batas terbesar dapat dirumuskan menjadi masalah optimasi konstrain, yaitu:

$$\frac{1}{2}|w|^2 \quad (2.5)$$

$$yi\ (xi.w + b) - 1 \geq 0 \quad (2.6)$$



Gambar 2.2 Hyperplane terbentuk diantara class-1 dan +1 (Susilowati et al., 2015)

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 kelas. *Pattern* yang paling dekat ini disebut sebagai *support vector*. Garis solid pada gambar menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada *Support Vector Machine* (Susilowati et al., 2015).

f. Kernel pada *Support Vector Machine*

Dalam kasus *machine learning*, *kernel trick* merupakan metode yang menggunakan algoritma

linier classifier untuk menyelesaikan permasalahan nonlinier dengan cara memetakkan dimensi input ke ruang dimensi yang lebih tinggi, sehingga membuat *linier classifier* di ruang dimensi yang baru setara dengan *nonlinear classifier* di ruang dimensi asli. Dengan kernel, fungsi pemetaan tidak pernah dihitung secara explisit, karena ruang dimensi tinggi yang digunakan memungkinkan pada dimensi yang tak terbatas (Pratama & Trilaksono, 2015). Berikut ini adalah beberapa fungsi kernel yang umum digunakan antara lain:

Tabel 2.3 Kernel pada Support Vector Machine

Kernel	Fungsi Kernel
Linier	x.y
Polynomial	$(x.y+c)^d$
RBF	$\exp(-\frac{\|\|x-y\|\|^2}{2\sigma^2})$
Sigmoid	tanh(k(x-y)+c)

g. MultiClass *Support Vector Machine*

Support Vector Machine saat pertama kali diperkenalkan oleh Vapnik, hanya dapat mengklasifikasikan data ke dalam dua kelas (klasifikasi biner). Namun, penelitian lebih lanjut untuk mengembangkan *Support Vector Machine* sehingga bisa mengklasifikasi data yang memiliki lebih dari dua kelas, terus dilakukan. Ada dua pilihan untuk mengimplementasikan *multi class Support Vector Machine* yaitu dengan

menggabungkan beberapa *Support Vector Machine* biner atau menggabungkan semua data yang terdiri dari beberapa kelas ke dalam sebuah bentuk permasalah optimasi. Namun, pada pendekatan yang kedua permasalahan optimasi yang harus diselesaikan jauh lebih rumit(Sembiring, 2007). Berikut adalah metode yang biasa digunakan dalam *MultiClass Support Vector Machine*

1. **Metode *One-Againts-All***

   Metode ini melakukan klasifikasi dengan membandingkan satu kelas dengan semua kelas lainnya kecuali kelasnya. Misalnya terdapat tiga kelas A, B, dan C maka perbandingan yang terbentuk adalah (A->B, C), (B->A, C), (C->A, B).

2. **Metode *One-Againts-One***

   Metode ini melakukan klasifikasi dengan membandingkan satu kelas dengan satu kelas lainnya, sehingga membentuk (k(k-1)/2) buah model klasifikasi biner (k adalah jumlah kelas). Misalnya terdapat tiga kelas A, B, dan C maka perbandingan yang terbentuk adalah (A->B), (B->C), (A->C).

3. **Metode DAGSVM (Directed Acyclic Graph Support Vector Machine)**

   Metode ini melakukan perbandingan sama seperti metode *One-Againts-One*, yaitu dengan membentuk (k(k-1)/2) buah model klasifikasi biner akan tetapi saat pengujian menggunakan

konsep *binary directed acyclic graph*. Konsep tersebut membentuk graph dan tiap kelas yang dilambangkan sebagai node. Pengunjian dimulai dari akar dan bergerak dari kiri ke kanan.

h. *Pre-Proccesing*

Preprocessing adalah salah satu langkah terpenting dari *Data Mining*. *Preprocessing* dilakukan untuk mendapatkan data yang akurat, Dalam *preprocessing* teks, ada banyak langkah seperti *Case Folding, Data Cleansing,* menghapus stopwords, *stemming* (Sharma, Agrawal, Lalit, & Garg, 2017).

1. *Case Folding* adalah proses dimana mengubah semua karakter pada teks menjadi huruf kecil dan menghilangkan angka atau bentuk ta.nda baca sehingga data yang didapat hanya mengandung karakter huruf a sampai z.
2. *Data Cleansing* adalah proses membersihan *tweet* dari kata yang tidak diperlukan atau untuk mengurangi *noise.*
3. Penghapusan *Stopwords* adalah proses menghilangkan kata-kata yang sering muncul tapi tidak memiliki makna dalam klasifikasi.
4. *Stemming* adalah proses menyederhanakan kata yang berisi imbuhan kembali ke kata dasarnya.

i. *SentiStrength*

SentiStrength adalah algoritma klasifikasi yang menggunakan pendekatan berbasis leksikon

yang menggunakan aturan-aturan dan informasi *linguistik* tambahan (non-leksikal) untuk mendeteksi kekuatan sentimen di sebuah teks singkat dalam bahasa Inggris. *SentiStrength* menggunakan sistem *dual-scale* (positif-negatif), karena menurut penelitian psikologi, manusia dapat merasakan emosi positif dan negatif secara bersamaan hingga batas tertentu secara mandiri. *SentiStrength* akan menghasilkan nilai positif dan negatif, dimana jangkauan nilai dimulai dari angka 1 sampai 5. Nilai 1 menunjukkan kalimat tersebut tidak memiliki sentimen positif maupun negatif. Kamus Sentimen. sedangkan nilai 5 menunjukkan kalimat tersebut memiliki sentimen sangat positif atau sangat negatif. Skor akhir pada sebuah kalimat ditentukan dari skor positif tertinggi dan skor negatif tertinggi dari kata-kata penyusunnya. Tabel 2.4 menunjukkan contoh dari hasil *SentiStrength*.

Tabel 2.4 Contoh hasil SentiStrength

No	Status
1	ahok kalah [-4] akibat jaring saracen wkwk
2	duh [-4] perih [-5] kalah [-5]
3	sukses [4] anies moga jd gubernur baik [4]

Angka di dalam tanda "[...]" menunjukkan skor kekuatan sentimen setiap kata penyusun kalimat yang sesuai dengan kamus sentimen kata.

Tabel 2.5 Menunjukkan aturan pembentukan keputusan akhir sentimen.

Tabel 2.5 Aturan dalam SentiStrength

No	Status
1	Nilai Positif > Nilai Negatif = Positif
2	Nilai Positif < Nilai Negatif = Negatif
3	Nilai Positif = Nilai Negatif = Netral

Selain menggunakan kamus sentimen dengan pembobotan term oleh manusia, *SentiStrength* juga menggunakan kamus emosikon dan kamus ungkapan yang juga diberi bobot oleh manusia. (Wahid & SN, 2016)

j. *K-fold Cross Validation*

K-fold Cross Validation adalah teknik untuk melakukan validasi pada dataset untuk menemukan akurasi yang baik. Teknik ini membagi dataset sebanyak *k subset*. Satu dari subset ini akan dijadikan sebagai data uji dan *k-1 subset* sisanya digunakan untuk proses data latih. Proses ini dilakukan sebanyak *k* kali sehingga setiap subset akan menjadi data uji dari model. Proses ini akan mendapatkan *k* buah nilai performa dari proses pembelajaran. Semua nilai performa ini akan dicari rata-ratanya dan nilai dengan rata-rata tertinggi akan dipilih sebagai model. *k-fold cross validation* memiliki kelebihan dapat mengklasifikasi *dataset* lebih efisien, namun metode ini memiliki kelemahan dalam proses komputasi yang digunakan akan

lebih besar karena akan melakukan proses sebanyak *k* kali (Haltuf, 2011).

k. N-Gram

N-gram adalah metode untuk mengurutkan kata yang sering muncul dalam sebuah teks. Secara umum, *n-gram* menciptakan ratusan atau ribuan variabel yang sering muncul dalam teks tertentu (Schonlau, Guenther, & Sucholutsky, 2017)