BAB II
TINJAUAN PUSTAKA

2.1 Kajian Pustaka

Kajian pustaka berisi penelitian terdahulu dengan topik relevan. Peneliti menggunakan sebelas jurnal topik analisis sentimen dengan komparasi metode SVM, NBC dan KNN. Tiga jurnal metode SVM, dua jurnal metode NBC, dua jurnal metode KNN, dua jurnal komparasi metode SVM dan NBC, dua jurnal komparasi metode SVM dan KNN.

Penelitian pada tahun 2016 meneliti tentang analisis sentimen terhadap *hatespeech* menggunakan dataset 522 *Tweet.* Objek penelitian berupa 522 *Tweet* terdiri dari 257 *hatespeech* dan 265 *goodspeech.* Metode penelitian melakukan komparasi terhadap algoritma SVM dengan NBC. Hasil penelitian adalah nilai *accuracy* SVM yang lebih tinggi dibanding NBC. Adapun nilai *accuracy* SVM sebesar 66,6% dan NBC sebesar 61,8% [8].

Penelitian yang kedua pada tahun 2019. Objek penelitian berupa ulasan produk beberapa situs jual beli *online.* Penelitian menggunakan tiga jenis data ulasan produk. Pertama 768 positif dan 564 negatif, kedua 768 positif dan 314 negatif, ketiga 268 positif dan 564 negatif. Metode klasifikasi yang digunakan adalah dengan melakukan komparasi antara algoritma SVM dan KNN. Kesimpulan dari hasil penelitian adalah akurasi terbaik di ketiga jenis data ulasan adalah dengan metode SVM. Perbandingan metode SVM dan KNN di ketiga jenis ulasan berturut-turut sebagai berikut, 88,83% dan 73,99% [14].

Penelitian selanjutnya menggunakan algoritma KNN sebagai metode klasifikasi untuk analisis sentimen. Penelitian ini menggunakan objek penelitian berupa *review* dari situs suatu agen travel. *Review* yang digunakan sebagai dataset sebanyak 200 yang terdiri 100 *review* dengan sentimen positif dan juga 100 *review* dengan sentiment negatif. Hasil dari penelitian menyimpulkan bahwa nilai *accuracy* sebesar 87% [10].

Penelitian berikut dilakukan pada tahun 2018 dengam membandingkan algoritma SVM dengan NBC untuk analisis sentimen. Objek penelitian berupa

1.491 *Tweet* yang dibagi menjadi dua kelas data yaitu 491 *Tweet* digunakan sebagai data latih dan 1.000 *Tweet* sebagai data uji. Hasil dari penelitian yang berupa komparasi algoritma SVM dengan NBC ini pun membuktikan bahwa nilai *accuracy* SVM lebih baik yaitu sebesar 81,67% sedangkan untuk NBC sebesar 67,20% [15].

Haranto pada tahun 2019 melakukan penelitian analisis sentimen menggunakan algoritma SVM. Objek penelitian berupa 500 *Tweet* dengan 250 *Tweet* mengenai Telkom dan 250 *Tweet* lainnya mengenai Biznet. Adapun hasil analisis sentimen dari penelitian ini untuk Telkom adalah 41,2% sentimen positif dan 58,8% sentimen negatif. Biznet pada penelitian ini memeperoleh 35,2% sentimen positif dan 64,8% sentimen negatif. Pada penelitian ini memiliki nilai *accuracy* sebesar 79,6% untuk Telkom dan 83,2% untuk Biznet [16].

Komparasi antara algoritma KNN dan SVM dalam analisis sentimen pernah diteliti pada tahun 2019. Dalam penelitian ini memiliki objek penelitian berupa 1.113 *Tweet* yang dibagi menjadi dua kelas data yaitu data latih dan data uji. Perbandingan data latih dan data uji adalah 80:20. Dari penelitian perbandingan antara KNN dengan SVM berikut diperoleh hasil bahwa nilai *accuracy* SVM lebih tinggi dibanding KNN. Nilai *accuracy* dari SVM sebesar 88,76% dan nilai *accuracy* dari KNN sebesar 88,1% [9].

Penelitian terdahulu berikut ini menggunakan metode NBC dalam melakukan analisis sentimen. *Dataset* yang digunakan sebagai objek penelitian sebanyak 221 *Tweet* yang terdiri dari 58 *Tweet* positif, 101 *Tweet* negatif, dan 62 *Tweet* netral. Hasil dari penelitian ini adalah nilai *accuracy* sebesar 81% pada data latih dan 76% pada data uji [17].

Pada tahun 2020 penelitian oleh Rian Tineges, *et. al.,* mengenai analisis sentimen menggunakan SVM. Data yang digunakan sebanyak 10.000 *Tweet* semenjak tanggal 16 hingga 23 Maret 2020 dan disimpan dalam CSV (*Comma Sparated Values).* Persentase penggunaan data latih dan datau uji adalah sebesar 80% data latih dan 20% data uji. Hasil dari penelitian didapat nilai *accuracy* sebesar 87%, nilai *precision* sebesar 86%, nilai *recall* sebesar 95% , nilai *error rate* sebesar 13%, dan nilai *F1-score* sebesar 90%. Hasil sentimen dengan metode SVM dapat

disimpulkan nilai sentimen positif sebesar 18,4% dan nilai sentimen negatif sebesar 81,6% [18].

Penelitian terdahulu yang selanjutnya adalah menggunakan NBC sebagai metode klasifikasinya. Penelitian yang dilakukan pada tahun 2021 ini menggunakan 12.906 *Tweet* sebagai dataset yang terdiri dari 8.942 *Tweet* terindikasi negatif, 3.830 positif dan 134 netral. Hasil dari penelitian analisis sentimen dengan menggunakan algoritma NBC ini memperoleh hasil berupa nilai *precission* sebesar 97,15% [19].

Penelitian yang juga dilakukan pada tahun 2021 berikut ini mengenai analisis sentimen menggunakan KNN. Objek dari penelitian adalah berupa *review* yang diambil dari aplikasi Google *Play Store*. *Review* yang digunakan sebanyak 4.362 yang selanjutnya dibagi menjadi dua kelas data yaitu 3.489 sebagai data latih dan 873 sebagai data uji. *Accuracy* yang dihasilkan penelitian adalah sebesar 81,70% [20].

Penelitian yang dilakukan pada tahun 2022 mengenai kombinasi metode SVM dan *Lexicon-Based* untuk penelitian dengan topik analisis sentimen. Objek yang digunakan pada penelitian adalah respons pengguna *Twitter* terhadap PP Tapera No. 25 Tahun 2020. Jumlah total *Tweet* yang digunakan sebanyak 519 *Tweet. Lexicon-Based Algorithm* digunakan untuk proses pelabelan sentimen pada data *Tweet.* SVM digunakan dalam penelitian ini sebagai metode klasifikasi. Perbandingan *acccuracy* SVM dilakukan dengan empat kernel yaitu linear, (*Radial Basis Function*) RBF, *Sigmoid* dan *Polynomial*. *Accuracy* terbaik dalam penelitian ini menggunakan kernel RBF sebesar 81,73% [21].

Empat dari sebelas jurnal terdahulu dalam kajian pustaka ini merupakan jurnal dengan komparasi algoritma untuk analisis sentimen. Dua jurnal diantaranya membandingkan SVM dengan NBC dan dua lainnya membandingkan SVM dengan KNN. Semua jurnal komparasi algoritma yang digunakan dalam kajian pustaka ini membuktikan bahwa *accuracy* dari algoritma SVM adalah yang paling tinggi. Berdasarkan dari kajian pustaka yang terangkum, penelitian ini menetapkan untuk menggunakan SVM.

Penelitian terdahulu selanjutnya disajikan dalam bentuk Tabel 2.1 Kajian Pustaka di bawah ini.

Tabel 2.1 Kajian Pustaka

No	Judul	Objek	Metode	Hasil	Perbedaan
1.	Analisis Sentimen *Hatespeech* pada *Twitter* dengan Metode *Naïve Bayes Classifier* dan *Support Vector Machine* [8]	Objek pada penelitian berupa 522 *Tweet*. 257 *Tweet* berupa *hatespeech* dan 265 *Tweet* berupa *goodspeech.*	Metode dengan menerapkan komparasi metode NBC dan SVM.	Kesimpulan dari hasil penelitian adalah nilai *accuracy* dengan algoritma SVM lebih baik yaitu 66,6% dan NBC sebesar 61,8%.	Penelitian Buntoro pada tahun 2016 melakukan komparasi metode untuk analisis sentimen antara algoritma NBC dan SVM.
2.	*Text Mining* dan Klasterisasi Sentimen pada Ulasan Produk Toko Online [10]	*Dataset* atau objek penelitian berupa ulasan produk di beberapa situs jual beli *online.* Penelitian menggunakan tiga jenis data ulasan produk.	Metode yang digunakan adalah KNN dan SVM untuk kemudian dilakukan komparasi.	Kesimpulan dari hasil penelitian adalah akurasi terbaik di ketiga jenis data ulasan adalah dengan metode SVM.	Penelitian ini melakukan komparasi antara metode SVM dan KNN dengan tiga jenis data ulasan. *Dataset* menggunakan ulasan produk pada situs jual beli.

No	Judul	Objek	Metode	Hasil	Perbedaan
3.	Penerapan Algoritma K-Nearest Neighbor pada Analisis Sentimen Review Agen Travel [14]	Objek penelitian diambil dari *review* atau komentar yang terdapat pada situs suatu agen travel. Dataset terdiri dari *review* positif dan negatif yang masing-masing terdapat 100 *review.*	Penelitian menggunakan metode KNN untuk melakukan analisis sentimen.	Hasil dari penelitian ialah nilai *accuracy* sebesar 87%.	Penelitian menggunakan algoritma KNN untuk analisis sentimen dan dataset berupa *review* dari suatu situs.
4.	Analisis Sentimen Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode *Support Vector Machine* dan *Naïve Bayes Classifier* [15]	Objek penelitian berupa 1.491 *Tweet*. Data dibagi menjadi 491 *Tweet* data latih dan 1.000 *Tweet* data uji.	Metode yang digunakan pada penelitian ini adalah NBC dan SVM untuk dilakukan sebuah komparasi.	Hasil penelitian menyimpulkan bahwa nilai *accuracy* SVM lebih tinggi yaitu sebesar 81,67%, sedangakan nilai *accuracy* NBC sebesar 67,20%.	Penelitian yang dilakukan pada tahun 2018 ini melakukan komparasi algoritma untuk analisis sentimen antara algoritma NBC dan SVM.

No	Judul	Objek	Metode	Hasil	Perbedaan
5.	Implementasi *Support Vector Machine* untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom dan Biznet [16]	Objek berupa 500 data *Tweet*. Sebanyak 250 adalah *Tweet* mengenai Telkom dan 250 data teks lainnya mengenai Biznet.	Penelitian menggunakan metode SVM untuk melakukan analisis sentimen.	Hasil penelitian menyimpulkan bahwa nilai *accuracy* sebesar 79,6% untuk Telkom dan *accuracy* 83,2% untuk Biznet.	Penelitian yang dilakukan Fadholi dan Bety pada tahun 2019 ini mencari tingkat kepuasan pelanggan terhadap pelayanan Telkom dan Biznet.
6.	Perbandingan Akurasi dan Waktu Proses Algoritma KNN dan SVM dalam Analisis Sentimen *Twitter* [9]	Objek penelitian berupa dataset yang berisi 1.113 data dan dibagi menjadi data latih dan data uji, dengan rasio berturut-turut adalah 80:20.	Metode yang diterapkan untuk sentimen analisis pada penelitian ini adalah KNN dan SVM.	Penelitian ini menarik kesimpulan bahwa SVM memiliki nilai *accuracy* yang lebih tinggi yaitu sebesar 88,76% dan KNN sebesar 88,1%.	Penelitian ini membandingkan metode KNN dan SVM untuk analisis sentimen.

No	Judul	Objek	Metode	Hasil	Perbedaan
7.	Analisis Sentimen Persepsi Masyarakat Terhadap Pemilu 2019 pada Media Sosial *Twitter* Menggunakan *Naïve Bayes* [17]	Dataset sebanyak 221 *Tweet* yang terdiri dari *Tweet* positif sebanyak 58, negatif 101, dan netral sebanyak 62 *Tweet.* Perbandingan rasio dari data latih dan data uji berturut-turut adalah 70:30.	Penelitian menggunakan metode NBC dalam penelitian analisis sentimen.	Hasil dari penelitian menunjukkan nilai *accuracy* dari data latih yaitu 81% sedangkan *accuracy* data uji sebesar 76%.	Penelitian yang dilakukan oleh Safitri Juanita pada tahun 2020 menggunakan metode NBC.
8.	Analisis Sentimen Terhadap Layanan Indihome Berdasarkan *Twitter* dengan Metode Klasifikasi *Support Vector Machine* (SVM) [18]	Objek penelitian sebanyak 10.000 *Tweet* dengan rasio penggunaan data latih dan data uji yaitu 80:20.	Metode menggunakan SVM dalam analisis sentimen.	Hasil penelitian dengan metode SVM didapat nilai *accuracy* 87%, *precision* 86%, *recall* 95%, *error rate* 13%, dan *F1-score* 90%.	Penelitian yang dilakukan pada tahun 2020 ini, melakukan analisis sentimen tentang opini pelanggan terhadap pelayanan Indihome.

No	Judul	Objek	Metode	Hasil	Perbedaan
9.	Analisis Sentimen Pembelajaran Daring pada *Twitter* di Masa Pandemi COVID-19 Menggunakan Metode *Naïve Bayes* [19]	Objek penelitian berupa 12.906 *Tweet.* Adapun komposisi dari 12.906 *Tweet* tersebut adalah 8942 *Tweet* yang diindikasi negatif, 3.830 positif dan 134 netral.	Metode dalam penelitian ini adalah metode NBC.	Hasil yang diketahui dari peneilitian ini adalah nilai *precission* sebesar 97,15%.	Penelitian ini menggunakan metode NBC untuk penelitian analisis sentimen terhadap pembelajaran daring.
10.	Analisis Sentimen pada Ulasan Penggunaan Bibit dan Bareksa dengan Algoritma KNN [20]	Objek penelitian adalah berupa *review* aplikasi pada Google Play Store. Dataset yang digunakan berjumlah 4.362 dengan 3.489 sebagai data latih dan 873 sebagai data uji.	Metode yang diterapkan pada penelitian analisis sentimen ini adalah dengan metode KNN.	Hasil dari penelitian dengan metode KNN ini mendapat nilai *accuracy* sebesar 81,70%.	Penelitian ini dilakukan oleh Aluisius dan Safitri pada tahun 2021 menggunakan metode KNN untuk analisis sentimen *review* pengguna suatu aplikasi.

No	Judul	Objek	Metode	Hasil	Perbedaan
11	*Combination of Support Vector Machine and Lexicon-Based Algorithm in Twitter Sentiment Analysis* [21]	Dataset penelitian berupa respons pengguna *Twitter* terhadap PP Tapera No. 25 Tahun 2020. Jumlah total *Tweet* yang digunakan sebanyak 519 *Tweet.*	Metode yang digunakan adalah SVM dan *Lexicon-Based* Algorithm	Hasil dari penelitian adalah *accuracy* tertinggi sebesar 81,73%.	Penelitian yang dilakukan pada April 2022 ini menggunakan dataset respons pengguna *Twitter* terhadap PP Tapera No. 25 Tahun 2020.

Sebagian besar penelitian ini mengacu kepada penelitian milik R. H. Muhammadi, dengan judul "*Combination of Support Vector Machine and Lexicon-Based Algorithm in Twitter Sentiment Analysis",* yaitu poin kesebelas pada Tabel 2.1 Kajian Pustaka. Dasar jurnal tersebut dijadikan acuan pada penelitian ini karena penelitian tersebut juga menggunakan *lexicon* pada proses *labeling*, dataset yang digunakan juga diperoleh dari data *tweet* serta pembagian data latih dan data uji sebesar 80:20. Adapun yang membedakan dengan penelitian milik R. H. Muhammadi adalah topik penelitian yang membahas tentang respons pengguna *Twitter* terhadap PP Tapera No. 25 Tahun 2020. Hal pembeda lainnya adalah jumlah kelas data pada penelitian milik R. H. Muhammadi menggunakan tiga kelas data yaitu negatif, netral dan positif, sedangkan pada penelitian ini menggunakan dua kelas data yaitu positif dan negatif [21].

2.2 Landasan Teori

2.2.1 Twitter

Twitter adalah media sosial yang pada awalnya berbasis *mobile* dengan cara mendaftar akun pada *Twitter*. Media sosial *Twitter* ini memampukan pengguna atau pemilik akun untuk dapat mengirim pesan berbentuk karakter atau teks tertulis secara *realtime* yang bersifat publik. Fungsi dari *Twitter* selain sebagai tempat berbagi pengalaman atau opini, juga dapat digunakan untuk pertukaran informasi dan berita dari seluruh penjuru dunia [22].

2.2.1.1 *Tweet*

Pesan karakter yang dapat diunggah melalui media sosial *Twitter* disebut dengan istilah *Tweet* [22] atau dalam bahasa Indonesia disebut kicauan [2]. *Tweet* tersebut memiliki keterbatasan karakter, *Twitter* membatasi *Tweet* hingga 140 karakter. *Tweet* yang diunggah pengguna bersifat publik, dapat dilihat siapapun. Pengguna yang saling mengikuti (*follow*) di *Twitter* akan dapat saling melihat *Tweet* yang mereka unggah, hal tersebut terjadi karena *Tweet* akan secara otomatis muncul di lini masa pengguna yang

mengikutinya (*follow*) [22].

2.2.1.2 *Mention*

Mention adalah istilah untuk memanggil pengguna *Twitter* lain dengan menyebut pengguna tersebut dalam *Tweet*. *Mention* dilakukan dengan penggunaan '@' lalu dilanjutkan dengan nama pengguna atau akun yang ingin di panggil [22].

2.2.1.3 *Hashtag*

Hashtag dapat disematkan pada sebuah *Tweet* untuk menandai sebuah bahasan atau topik pembicaraan yang sama di *Twitter*. Cara menggunakan *hashtag* adalah dengan mengetikkan '#' lalu diikuti dengan topik yang ingin dibahas [22].

2.2.1.4 *Trending Topics*

Trending topics pada *Twitter* berisi kumpulan topik atau *hashtag* yang paling populer. *Hashtag* pada *Twitter* terakumulasi dan dilihat berdasarkan jumlah *Tweet* untuk topik tersebut, biasanya topik yang paling banyak dibahas yang akan ditampilkan [22].

2.2.2 Jupyter Notebook

Jupyter Notebook seringkali digunakan sebagai catatan dari alur program. Kode program dapat disunting atau dimodifikasi untuk kemudian dijalankan kembali. *Jupyter Notebook* memiliki fungsi utama sebagai wadah pencatat sintaks yang interaktif berbasis web lengkap dengan fitur visualisasi [23].

2.2.3 Python

Bahasa pemrograman *python* banyak digunakan *programmer* karena mudah dipelajari dan struktur sintaks yang baik juga mudah dipahami. Pemanfaatan *python* yang paling populer adalah untuk keperluan website dan internet, penelitian ilmiah dan numerik, data science dan big data, media pembelajaran pemrograman, *Graphical User Interface* (GUI), pengembangan software dan aplikasi bisnis. Bahasa pemrograman *python* dikenal sebagai bahasa pemrograman yang dinamis dengan komunitas yang besar dan *python* juga memiliki pustaka (*library*) umum yang lengkap [24].

Library pada *python* lengkap untuk memproses suatu kasus *text mining.*

Adapun *library* yang digunakan pada penelitian ini adalah *pandas* untuk mengambil dan juga membaca data, *tweepy* untuk pengambilan data *Tweet* pada *Twitter* API, *matplotlib* untuk visualisasi data, re, nltk, dan *collections* untuk tahapan *text mining* [25].

2.2.4 Kecerdasan Buatan

Kecerdasan buatan adalah implementasi dari kecerdasan yang ditanam pada mesin atau software agar dapat berpikir seperti manusia. Kecerdasan buatan banyak dimanfaatkan untuk memecahkan masalah beberapa bidang. Contoh bidang yang menggunakan kecerdasan buatan adalah robotika, diagnosa medis, persepsi, dan lain-lain [26].

2.2.5 Machine Learning

Machine learning merupakan cabang dari kecerdasan buatan. *Machine learning* adalah aplikasi komputer dengan algoritma matematika agar dapat melakukan pembelajaran secara mandiri. Pembelajaran yang dilakukan oleh mesin berdasarkan data yang diberikan dan akan hasil dari *machine learning* berupa prediksi. *Machine learning* terbagi menjadi tiga kategori yaitu, *supervised learning*, *unsupervised learning*, dan *reinforcement learning* [26]. *Supervised learning* adalah metode dengan memberikan data latih kepada sistem yaitu data yang sudah memiliki output atau label untuk kemudian dipelajari oleh sistem, sedangkan untuk *unsupervised learning* adalah metode tanpa adanya pelatihan (*trainning*) sistem dengan data [27]. *Reinforcement learning* adalah metode diantara *supervised* dan *unsupervised learning* yang ditujukan untuk dapat bekerja dalam lingkungan yang dinamis, teknik yang digunakan ialah *trial and error* dengan mempelajari langsung dari pengalaman barunya [26].

2.2.6 Text Mining

Text mining merupakan proses interaksi penggunanya dengan koleksi dokumen di rentang waktu tertentu. Dokumen yang berbentuk *text* tersebut dikumpulkan dengan beberapa alat dan proses analisis [20]. Proses yang dialami dalam *text mining* secara garis besar ada dua, yang pertama

penerapan struktur. Penerapan struktur dilakukan untuk membuat data *text* yang sebelumnya dianggap sebagai data tidak terstruktur diubah menjadi data terstruktur. Proses yang kedua ialah ekstraksi informasi yang relevan dari data *text* yang dianalisa [4].

2.2.7 Analisis Sentimen

Analisis sentimen atau yang biasa juga disebut *opinion mining* merupakan cabang ilmu dari *text mining* [16]. *Opinion mining* dapat melakukan analisis suatu sentimen, sikap atau emosi seseorang terhadap suatu topik seperti layanan, politik, peristiwa, dan lain-lain. Analisis sentimen yang juga dikenal sebagai *opinion mining* berfokus untuk mengetahui kecenderungan sentimen terhadap suatu topik yang dapat bersifat postif maupun negatif [22]. Adapun langkah umum dalam proses analisis sentimen adalah sebagai berikut [10]:

2.2.7.1 Pengumpulan *Dataset*

Proses yang pertama kali dilakukan ialah pengumpulan dataset. Dataset dapat berupa *review*, komentar, *Tweet*, dan lain-lain dari suatu aplikasi atau website [10].

2.2.7.2 *Pre-processing*

Pada tahap ini proses yang dilakukan ialah membersihkan data dengan menyamaratakan data sesuai standar agar dapat terbaca dengan baik oleh sistem [10]. Tahap *pre-processing* biasanya memiliki tahapan berupa *case folding*, *filtering*, *tokenizing*, *stopwards removal*, dan *stemming*. *Case folding* adalah penyamarataan huruf atau karakter menjadi huruf kecil (*lowercase*) semua. *Filtering* adalah pembersihan data dari hal-hal yang tidak diperlukan dalam proses analisis sentimen seperti tanda baca, *Uniform Resource Locator* (URL), dan lain-lain. *Tokenizing* adalah untuk mengubah kalimat menjadi kata terpisah. *Stopward removal* adalah proses menghapus kata sambung dan kata ganti orang agar kata kunci yang akan diproses selanjutnya dalam pembobotan lebih

sempit [18]. *Stemming* yaitu mengembalikan kata berimbuhan ke kata dasar [15].

2.2.7.3 Pembobotan TF-IDF

Pembobotan atau pemberian nilai pada data teks. Proses yang sering digunakan pada tahap ini adalah TF-IDF [10]. TF-IDF merupakan proses pembobotan kata dimana kepanjangan dari TF adalah *Term Frequency* yang merupakan frekuensi dari tingkat kemunculan suatu kata dalam sebuah dokumen. IDF memiliki kepanjangan *Inverse Document Frequency* yaitu maksud dari *Document Frequency* sendiri adalah jumlah dokumen yang memiliki kemunculan dari kata tersebut [28]. Adapun rumus dari TF-IDF adalah sebagai berikut:

$$\text{TF-IDF} = \text{TF}\ (log\ \frac{\text{d}}{\text{df}}) \qquad (2.1)$$

Keterangan:

TF = Jumlah kemunculan kata dalam satu dokumen

IDF = *Inverse Document Frquency*

d = Jumlah dokumen

df = Jumlah dokumen yang terdapat kata (yang sedang dicari)

Proses pertama yang dilakukan adalah mencari TF dengan menghitung jumlah kemunculan setiap *term* pada masing-masing dokumen. Kedua, mencari DF dengan menghitung jumlah dokumen terhadap setiap *term.* Langkah selanjutnya ketika nilai dari TF dan IDF sudah didapatkan, jumlah kemunculan setiap *term* dalam masing-masing dokumen dikali dengan hasil IDF setiap term [29].

2.2.7.4 *Classification*

Pada tahap ini akan dilakukan proses pengklasifikasian sentimen yang biasanya menggunakan metode seperti NBC, KNN, SVM, dan lain-lain [10].

2.2.7.5 *Evaluation*

Tahap *evaluation* adalah tahap yang memproses perhitungan *accuracy, precission, recall* dan *f1-score* [20]. *Accuracy* adalah persentase dari total ketepatan prediksi sentimen. *Precission* adalah persentase dari ketepatan prediksi terhadap sentimen positif dengan total prediksi terhadap sentimen positif. *Recall* persentase dari ketepatan prediksi terhadap sentimen positif dengan total data aktual positif [30].

2.2.8 Lexicon

Lexicon adalah metode pelabelan kalimat ke dalam sentimen negatif atau positif. Konsep dari *lexicon* yaitu penggunaan kamus yang berisi kata positif dan negatif sebagai bahan pembanding dalam menghitung tingkat polaritas suatu kalimat untuk dilabeli sebagai sentimen negatif atau positif. Terdapat dua jenis kamus pada *lexicon,* yaitu kamus dengan nilai polaritas pada setiap kata dan kamus yang nilai atau skor pada kata bergantung dari tingkat polaritasnya [31].

2.2.9 Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan salah satu dari beberapa metode klasifikasi dengan teknik *supervised* dan memiliki nilai *accuracy* yang baik [15]. Nilai dari *accuracy* SVM yang baik dipengaruhi karena SVM adalah metode yang dapat mengenali persebaran pola kata di suatu kalimat dengan *hyperplane* [8]. Konsep dari SVM menggunakan *hyperplane* dan menemukan *hyperplane* terbaik dari setiap kemungkinan [9]. *Hyperplane* sendiri merupakan garis pemisah atau pembatas yang mengklasifikasi menurut kelasnya. Garis pemisah yang baik adalah jika garis memiliki jarak terbesar terhadap data latih terdekat di setiap kelasnya. Semakin besar margin, maka *error* generalisasinya juga akan semakin rendah, maksud dari margin sendiri ialah jarak yang dihitung dari suatu titik vektor yang berada didalam kelas terhadap *hyperplane* [16].

SVM secara garis besar adalah mengubah data yang sebelumnya berupa *text* menjadi *vector* data yang juga dikombinasikan dengan hasil dari

pembobotan yaitu TF-IDF. Fungsi dari persamaan SVM dalam persamaan (2.2) di bawah ini:

$$h(X) = z^{x} \varnothing(X) + c \tag{2.2}$$

Keterangan:

h (X) = fungsi SVM

z = fitur *vector* (dari bobot yang berbeda)

X = fitur *vector*

Ø = fungsi pemetaan non linier

c = *vector* bias

Dalam persamaan (2.2) diatas, X adalah fitur *vector* dan z juga *vector,* namun berasal dari bobot yang berbeda. Selanjutnya Ø merupakan fungsi dari pemetaar non-linier. C dalam persamaan (2.2) merupakan *vector* bias, dan biasanya z dan c dapat diperoleh dari data latih secara otomatis [15].

2.2.10 Confussion Matrix

Confusion matrix dapat memberikan keputusan yang diperoleh dari proses menggunakan data latih dan data uji. Tahap ini juga memberi penilaian terhadap kinerja dari metode dalam menentukan klasifikasi [32]. Fungsi lain dari *Confusion matrix* adalah untuk mengetahui nilai dari *accuracy, precission, recall,* dan *f1-score* [16]. *Confusion matrix* digunakan sebagai tolak ukur seberapa baik proses dari pengkalsifikasian mengenali tiap *Tweet* dari sentimen yang berbeda [32]. *Confusion matrix* seperti Tabel 3.1 di bawah ini.

Tabel 3.1 Confusion Matrix

Aktual	Prediksi	
	True	False
True	TP	FN
False	FP	TN

Parameter TP (*True Positive)* ada ketika kondisi prediksi positif dengan aktual yang juga positif. Parameter FN (*False Negative*) ketika kondisi prediksi negatif sedangkan aktualnya positif. Parameter FP (*False Positive*) ketika kondisi prediksi positif sedangkan aktualnya negatif. Parameter TN (*True Negative*) ketika kondisi prediksi negatif dan aktual negatif. Perhitungan untuk mendapatkan nilai *accuracy, precision, dan recall* berkaitan dengan *confusion matrix* seperti pada persamaan (3.2) hingga persamaan (3.4) di bawah ini [29].

$$A = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{3.2}$$

$$P = \frac{(TP)}{(TP + FP)} \tag{3.3}$$

$$R = \frac{(TP)}{(TP + FN)} \tag{3.4}$$

F-measure adalah berupa *harmonic mean* dari nilai *precission* dan *recall.* Perhitungan untuk *f1-score* seperti pada persamaan (3.5) di bawah ini [9].

$$f1 - score = \frac{(2 \times P \times R)}{(P + R)} \tag{3.5}$$

Keterangan:

A = *Accuracy*

P = *Precission*

R = *Recall*

F1-score = f1-score

TP = jumlah prediksi positif dan aktual positif

TN = jumlah prediksi negatif dan aktual negatif

FP = jumlah prediksi positif dan aktual negatif

FN = jumlah prediksi negatif dan aktual positif