BAB II
TINJAUAN PUSTAKA

Bab ini berisi tentang tinjauan pustaka yang berhubungan dengan hasil penelitian terdahulu yang dilakukan oleh peneliti sebelumnya serta berisi penjabaran teori yang berkaitan dengan penelitian ini. Hal tersebut digunakan sebagai dasar membuat struktur landasan teori.

2.1 Penelitian terdahulu

Penelitian pertama berjudul "*Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches*" yang ditulis oleh Heidi Nguyen, Aravinnd Veluchamy, Mamadou Diop dan Moh. Rashed Iqbal pada tahun 2018. Penelitian ini bertujuan untuk menganalisis sentimen dari review sejumlah produk yang ada di Amazon. Data yang digunakan didapat dari hasil *crawling* twitter menggunakan *library* python yaitu *Scapy*. Penelitian ini menggunakan dua pendekatan analisis sentimen, yaitu dengan *Lexicon-Based* dan *Machine learning*. Untuk analisis yang berdasarkan *lexicon,* metode yang digunakan adalah VADER, SentiWordNet dan *Pattern lexicon.* Sedangkan untuk pendekatan *machine learning,* algoritma yang digunakan adalah SVM, *Logistic Regression*, dan *Gradient Boosting*. Data yang akan dianalisis terlebih dahulu dilakukan *preprocessing* yang meliputi *parsing*, menghilangkan html, *stopwords removal, special character removal, lemmatization*, dan *tokenization.* Dalam penelitian ini juga menggunakan TF-IDF sebagai *feature-extraction*. Hasil klasifikasi diperoleh untuk pendekatan *Lexicon-based* yang menggunakan model Pattern, VADER, and SentiWordNet memiliki akurasi sebesar 69%, 83%, dan 80%. Dan untuk pendekatan *machine learning*, hasil akurasi yang diperoleh dengan algoritma SVM, *Gradient Boosting,* dan LR model adalah sebesar 89%, 87%, and 90%. Dari hasil klasifikasi yang ditunjukkan dalam penelitian ini, Pendekatan berbasis *machine learning* memiliki hasil yang lebih baik dibandingkan hasil klasifikasi menggunakan *Lexicon-based* [17].

Penelitian kedua di tahun 2019 berjudul "Analisis Sentimen Pengguna Gopay Menggunakan Metode *Lexicon-Based* dan *Support Vector Machine*" yang ditulis oleh Rachmad Mahendrajaya, Ghulam Asrofi Buntoro, dan Moh. Bhanu Setyawan. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap aplikasi dompet digital Gopay di media sosial twitter. Data yang digunakan didapat dari hasil *crawling* twitter menggunakan Twitter API dengan mencari *keyword* "Gopay" dari tanggal 24 Juli 2019 sampai 30 Juli 2019. Data yang terkumpul berjumlah 1210 tweet yang berbahasa Indonesia. Penelitian ini menggunakan algoritma *Support Vector Machine* dan pelabelan data yang menggunakan metode *Lexicon-based*. Hasil evaluasi menunjukkan bahwa algoritma *Support Vector Machine* memiliki tingkat akurasi sebesar 89,17% untuk kernel linear dan 84,38% untuk kernel polynomial. Pada algoritma *Support Vector Machine*, kernel linear menghasilkan tingkat akurasi yang lebih baik dibandingkan dengan kernel polynomial. Saran yang diberikan oleh peneliti untuk penelitian selanjutnya adalah untuk memperbanyak data dan menambah algoritma pembanding lainya seperti Naive Bayes untuk variasi penelitian [18].

Penelitian ketiga berjudul "Analisis Sentimen Opini Publik Tentang Undang-Undang Cipta Kerja Pada Twitter" yang ditulis oleh Tamora Nonia Wijaya, Rini Indriati, dan Muhammad Najibulloh Muzaki di tahun 2021. Penelitian ini bertujuan untuk menganalisis sentimen publik terkait dengan Undang-undang Cipta Kerja dari media sosial twitter. Peneliti menggunakan algoritma Naive Bayes dengan pembobotan TF-IDF. Data yang digunakan dalam penelitian ini berjumlah 1000 tweet dengan periode tweet yang dikirim dari Oktober 12020 sampai November 2020. Data selanjutnya diseleksi untuk mendapatkan data yang mengandung opini. Hasil seleksi adalah sebanyak 687 yang kemudian melewati tahap *preprocessing* yang meliputi *cleaning*, *case folding*, *tokenization*, normalisasi, *filtering*, dan *stemming*. Kemudian dilakukan proses *text transformation* menggunakan TF-IDF. Setelahnya dilakukan pemodelan dengan menggunakan algoritma Naive Bayes. Dari hasil pengujian yang dilakukan, diperoleh bahwa performa Naive Bayes dengan pembobotan TF-IDF memiliki nilai akurasi 89,9% [19].

Penelitian selanjutnya yaitu penelitian yang dilakukan di tahun 2021 oleh Hashri Hayati, dan Muhammad Riza Alifi dengan judul "Analisis Sentimen Pada Tweet Terkait Vaksin Covid-19 Menggunakan Metode *Support Vector Machine"*. Penelitian ini bertujuan untuk menganalisis opini masyarakat terhadap vaksin Covid-19. Metode yang digunakan dalam penelitian ini adalah *Support Vector Machine*. Data yang digunakan berupa tweet yang berkaitan dengan formula kata kunci yang terdiri dari merek vaksin, sinonim dari kata vaksin, dan organisasi pembuat vaksin dalam rentang waktu tiga bulan terakhir, yaitu bulan Maret 2021 sampai Mei 2021 Sebelum dimasukkan dalam model klasifikasi, data terlebih dahulu melewati proses *preprocessing* yang meliputi normalisasi kata, anotasi kata, *unpack hashtag*, *unpack contraction*, dan *transform emoticon*. Ada tiga metode *feature extraction* yang digunakan, yaitu TF-IDF, bigram, dan unigram. Hasil klasifikasi didapatkan bahwa metode *Support Vector Machine* dengan tokenisasi unigram dan bigram tidak memiliki perbedaan yang signifikan, yaitu selisih hanya 0.6% - 0.7%. Hasil akurasi tertinggi diperoleh sebesar 84% [20].

Dalam penelitian yang lain di tahun 2021, dilakukan perbandingan antara metode Naïve Bayes dan *Support Vector Machine*. Penelitian yang dilakukan oleh Brian Laurensz dan Eko Sediyono bertujuan untuk menganalisis tanggapan masyarakat terhadap tindakan vaksinasi di media sosial twitter. Data yang digunakan dalam penelitian ini berjumlah sebanyak 845 tweet yang diambil dari twitter menggunakan teknik *webscraping* dengan *tools* Octoparse berdasarkan dua kata kunci, yaitu 'vaksinmerahputih' dan 'vaksinsinovac'. Beberapa tahap pre-processing dalam penelitian ini adalah *transform cases, tokenize, filter token, stopword removal, stemming*, dan *Generate n-Grams*. Hasil dari penelitian ini menunjukkan bahwa penggunaan algoritma Naïve Bayes dengan validasi 10 *k-fold cross validation* mempunyai nilai akurasi terbaik dibandingkan metode *Support Vector Machine* dengan tingkat persentase akurasi sebesar 85,59%. Saran dari penelitian ini adalah agar penelitian selanjutnya melakukan *preprocessing* yang lebih baik lagi dan menambah sumber data yang lain [21].

Telah dilakukan juga penelitian di tahun 2022 juga yang membahas tentang analisis sentimen mengenai vaksin yang berjenis Sinovac dengan data yang berasal dari media sosial twitter. Dalam penelitian ini, data diambil menggunakan metode *crawling* dengan python selama 6 jam pada setiap kata kunci. Proses klasifikasi data dalam penelitian ini menggunakan metode Naïve Bayes yang dikategorikan menjadi 2 jenis yaitu positif dan negatif. Hasil dari penelitian ini menunjukkan bahwa Naïve Bayes mampu mengklasifikasikan tweet dan mengolompokkannya dalam sentimen positif dan negatif. Nilai probabilitas yang dihasilkan adalah 0,000002765 untuk positif dan 0,000000359 untuk negatif [22].

Berdasarkan penjelasan dari beberapa penelitian sebelumnya, maka ringkasan penelitian-penelitian tersebut dapat dimuat didalam tabel 2.1 berikut:

Tabel 2.1 Penelitian Terdahulu Terkait Analisis Sentimen

No	Judul	Objek	Metode	Hasil	Perbedaan
1	*Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches* (2018) [17]	Review 1000 produk di Amazon	Lexicon-based (VADER, Pattern, SentiWordNet) dan *Machine Learning* (*Logistic Regression*, *Support Vector Machine*, dan *Gradient Boosting*)	Hasil dari penelitian ini menunjukkan bahwa metode yang menggunakan *machine learning* memiliki akurasi yang lebih baik dengan hasil sebesar 89%, 87%, and 90% untuk SVM, GB, dan LR. Sedangkan sentimen menggunakan *lexicon-based* memperoleh hasil 69%, 83%, dan 80% untuk Pattern, VADER, dan SentiWordNet	Pada penelitian tersebut dilakukan perbandingan metode *Lexicon-based* dan *Machine learning* sebagai metode untuk klasifikasi data sedangkan pada penelitian ini *Lexicon-based* digunakan untuk pelabelan data sedangkan untuk klasfikasi data menggunakan pendekatan *Machine Learning* dengan algoritma Naïve Bayes dan SVM

No	Judul	Objek	Metode	Hasil	Perbedaan
2	Analisis Sentimen Pengguna Gopay Menggunakan Metode *Lexicon Based* dan *Support Vector Machine* (2019) [18]	Tweet pada twitter tentang Gopay berjumlah 1210	Lexicon Based dan *Support Vector Machine*	Hasil menunjukkan bahwa metode *Support Vector Machine* dengan pelabelan menggunakan *Lexicon-Based* dapat melakukan klasifikasi sentimen dengan akurasi sebesar 89.17%	Pada penelitian tersebut data yang dipakai merupakan data dengan ulasan pelanggan terkait dompet digital Gopay sedangkan pada penelitian ini data yang akan dianalisis adalah tentang Vaksin Covid-19 jenis Astrazeneca.
3.	Analisis Sentimen opini publik tentang undang-Undang Cipta Kerja Pada Twitter (2021) [19]	Tweet pada twitter tentang UU Cipta Kerja	Naïve Bayes	Hasil dari penelitian ini menunjukkan bahwa Naïve Bayes memiliki performa akurasi sebesar 89.9%. Dengan hasil sentimen 52.9% negatif dan 47.1% positif	Pada penelitian tersebut data yang digunakan adalah opini public terkait dengan UU Cipta Kerja sedangkan pada penelitian ini adalah tentang Vaksin jenis Astrazeneca
4	Analisis Sentimen Pada Tweet Terkait Vaksin Covid-19 Menggunakan Metode *Support Vector Machine* (2021) [20]	Tweet pada twitter dengan beberapa *keyword* yang telah ditentukan pada formula dalam penelitian ini	*Support Vector Machine*	Hasil evaluasi menggunakan tokenisasi unigram dan bi-gram memperoleh nilai akurasi tertinggi sebesar 84%	Pada penelitian tersebut, metode yang digunakan adalah *Support Vector Machine*, sedangkan pada penelitian ini metode yang digunakan adalah Naïve Bayes dan *Support Vector Machine*

No	Judul	Objek	Metode	Hasil	Perbedaan
5	Analisis Sentimen Masyarakat terhadap Tindakan Vaksinasi dalam Upaya Mengatasi Pandemi Covid-19 (2021) [21]	Tweet pada twitter yang berjumlah sebanyak 845 tweet dengan kata kunci "vaksin sinovac" dan "vaksin merah putih"	Naïve Bayes dan *Support Vector Machine*	Penelitian ini menunjukkan bahwa hasil klasifikasi dengan menggunakan metode Naïve Bayes mempunyai ratarata tingkat akurasi lebih besar dengan persentase sebesar 85,59%, sedangkan metode SVM 84,41%.	Pada penelitian tersebut dilakukan penelitian terhadap vaksin Covid-19 secara umum, sedangkan pada penelitian ini berfokus pada jenis vaksin Astrazeneca. Pada proses pelabelannya juga masih dilakukan manual. Sedangkan pada penelitian ini menggunakan pelabelan menggunakan *lexicon-based*
6	Analisis Sentimen Mengenai Vaksin Sinovac di Media Sosial Twitter Menggunakan Metode Naïve Bayes Classification (2022) [22]	Tweet pada twitter tentang Vaksin Sinovac	Naïve Bayes	Hasil dari penelitian ini menunjukkan bahwa probabilitas akhir berdasarkan kondisi 0,000002765 untuk positif dan 0,000000359 untuk negatif.	Pada penelitian tersebut, data yang digunakan berfokus pada vaksin jenis Sinovac sedangkan pada penelitian ini akan berfokus ke jenis Astrazeneca.

2.2 Dasar Teori

2.2.1 *Coronavirus Disease* 2019 (COVID-19)

Coronavirus atau (SARS-CoV-2) merupakan virus RNA *strain* tunggal positif, yang berkapsul dan tidak bersegmen. Virus ini dapat memasuki saluran napas yang kemudian bereplikasi di sel epitel saluran napas dan juga dapat menular melalui cairan dari hidung ataupun air liur saat yang terinfeksi bersin atau batuk. Masa inkubasi virus dari pertama memasuki sel sampai muncul penyakit adalah sekitar 3 hingga 7 hari. Infeksi dari COVID-19 dapat menimbulkan gejala ringan, sedang hingga gejala berat. Gejala klinis yang paling umum adalah demam, kelelahan, dan gejala saluran napas lain. Pada kasus yang berat, virus ini dapat menyebabkan infeksi saluran napas, distress pernapasan akut, dan penurunan saturasi oksigen hingga kurang dari 90%. Kunci pencegahan dari COVID-19 adalah untuk memutus rantai penularan adalah dengan isolasi, deteksi dini, melakukan proteksi dasar, serta membuat imunitas untuk mencegah transmisi dengan cara vaksin [23].

2.2.2 Vaksinasi COVID-19

Vaksinasi adalah proses dimana seseorang memperoleh kekebalan atau terlindungi dari suatu penyakit sehingga apabila suatu hari terinfeksi oleh penyakit tersebut maka tidak mengalami sakit atau hanya sakit ringan, biasanya dengan pemberian vaksin. Vaksinasi COVID-19 merupakan bagian dari upaya penanganan pandemi COVID-19 yang mencakup aspek preventif dengan penerapan protokol kesehatan: menjaga jarak, mencuci tangan pakai sabun dan memakai masker (3M), vaksinasi COVID-19, dan 3T (Tes, Telusur, Tindak lanjut).

Tujuan utama dari vaksinasi COVID-19 adalah untuk mengurangi infeksi/penularan COVID-19 dan menurunkan angka infeksi dan kematian akibat COVID-19 dan mencapai kekebalan kelompok *(herd immunity)*, melindungi masyarakat dari COVID-19 serta menjaga produktivitas sosial

dan ekonomi. *Herd Immunity* atau kekebalan kelompok terbentuk ketika mayoritas populasi divaksinasi. Cakupan vaksinasi yang tinggi membutuhkan partisipasi dan kerjasama berbagai pihak untuk mengatasi keengganan dan keraguan *(hesitancy)* masyarakat terhadap vaksinasi, meningkatkan penerimaan *(acceptance)* dengan memastikan ketersediaan akses informasi yang akurat tentang vaksinasi COVID-19. [2].

2.2.3 Vaksin Astrazeneca

Vaksin Oxford–Astrazeneca COVID-19 (AZD1222) merupakan vaksin berbasis protein yang mengekspresikan adenovirus (ChAdOx1) yang dimodifikasi dari simpanse. Di dalamnya terdapat urutan genetik yang dikodekan oleh SARS-CoV-2, yang mengarahkan sel untuk menghasilkan protein lonjakan COVID-19. Hal ini yang menyebabkan respons imun merangsang tubuh untuk serangan virus di masa depan. Dalam fase 2/3 uji klinis vaksin Oxford-Astrazeneca, nyeri ketika tekan di tempat suntikan adalah reaksi lokal yang paling sering dilaporkan, dengan kelelahan, sakit kepala, demam, dan mialgia sebagai efek samping sistemik yang paling umum [4].

Keunggulan vaksin Astrazeneca adalah dapat mencegah COVID-19 yang parah pada potensi lansia, dan pada semua orang dewasa. Vaksin jenis ini harus disimpan pada suhu 2-8 derajat celsius dan dapat digunakan dalam waktu 6 jam setelah botol dibuka. Suntikan diberikan dengan cara injeksi intrasmuskular (jaringan otot) kepada sasaran subjek yang berusia diatas 18 tahun. Vaksinasi diberikan dua kali dengan dosis dengan 0.5 ml, dengan interval injeksi 8-12 minggu [24].

2.2.4 Polaritas Sentimen

Polaritas sentimen adalah titik pada skala evaluasi yang sesuai dengan evaluasi positif atau negatif tentang sentimen tersebut. Menurut Kamus Besar Bahasa Indonesia, sentimen positif adalah reaksi atau sikap setuju, sependapat yang dapat menambah nilai pada seseorang atau sesuatu.

Di sisi lain, sentimen negatif menurut KBBI adalah reaksi yang merendahkan nilai dari seseorang atau sesuatu sehingga mengurangi *trend* hal tersebut. Secara umum, frasa dengan sentimen negatif ditandai dengan penggunaan kata negasi. Polaritas sentimen dapat bersifat positif, seperti 'Permainan yang ada di ponsel ini sangat bagus! Saya sangat menikmatinya!', atau negatif, seperti 'Saya merasa sedih dengan ponsel dari merk ini, karena tidak bisa hidup lagi' [25].

2.2.5 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah bidang ilmu kecerdasan buatan yang menggunakan bahasa alami untuk mempelajari komunikasi antara manusia dengan komputer. NLP menganalisis dan merepresentasikan teks yang tertulis secara alami (menggunakan bahasa manusia) pada satu atau lebih tingkat analisis linguistik dengan tujuan untuk mendapatkan pemrosesan bahasa mirip manusia yang dapat diimplementasikan dalam berbagai bidang.

Dalam implementasinya, NLP mempunyai beberapa tantangan untuk memahami bahasa manusia, diantaranya adalah sebagai berikut :

a. *Part of speech tagging* (Penandaan kelas kata)
   Part of speech tergantung pada konteks kata, sehingga sulit untuk mengidentifikasi spesifikasi dari *part of speech* (kata benda, kata kerja, kata sifat, dan sebagainya) dalam suatu teks.
b. *Text Segmentation* (Segmentasi teks)
   Penentuan segmentasi sulit dibuat dalam bahasa tertulis tanpa pembatas kata khusus. Misalnya dalam bahasa Mandarin, Jepang, dan Thailand dan pada bahasa lisan dimana suara dapat membaur diantara kata-kata.
c. *Word sense disambiguation* (Disambiguasi makna kata)
   Sebuah kata yang memiliki banyak arti baik dalam bentuk homonim (makna berbeda dan arti yang berbeda), misalnya perbedaan antara arti dari "bisa" yang dapat bermakna "dapat" dan "racun") maupun kata

polisemi (makna yang berbeda, tetapi terkait, contohnya "ragu" dalam makna "bimbang" dan "sangsi"). Pembeda makna adalah konteks dari penggunaanya yang dilakukan dengan mempertimbangkan.

d. *Syntatic ambiguity* (Ambiguitas sintaksis)

   Bahasa memiliki struktur kalimat yang berbeda. Kombinasi informasi semantik dan kontekstual biasanya diperlukan untuk memilih struktur yang paling tepat.

e. *Imperfect or irregular* input (Masukan yang tak sempurna atau tak teratur)

   Aksen yang diucapkan dan kesalahan ejaan dan tata bahasa dalam bahasa tulis sehingga dapat membuat pemrosesan bahasa alami menjadi sulit.

f. *Speech act* (Pertuturan)

   Hanya menggunakan struktur kalimat saja tidak secara akurat menjelaskan maksud dari pembicara atau penulis. Dari waktu ke waktu, gaya bahasa dan konteks menentukan makna yang dimaksudkan [26].

2.2.6 Analisis Sentimen

Analisis sentimen adalah proses memahami dan memproses, dan mengekstrak informasi secara otomatis dari data tekstual. Analisis sentimen biasanya dilakukan untuk menganalisis data tekstual berupa opini tentang subjek dan objek dalam kumpulan data yang mengandung polaritas sentimen. Saat ini, para peneliti sering menggunakan analisis sentimen sebagai cabang penelitian dalam ilmu komputer yang bertujuan untuk memberikan informasi dari kumpulan data yang tidak terstruktur [27].

2.2.7 Twitter

Twitter merupakan salah satu media sosial yang paling populer yang bertindak sebagai sebagai platform komunikasi sosial. Twitter memungkinkan orang di seluruh dunia untuk terhubung dengan keluarga, teman, dan kerabat menggunakan komputer dan ponsel mereka. Salah satu

layanan yang disediakan oleh twitter kepada pengguna adalah pembuatan pesan status (disebut tweet) yang dapat dibaca oleh pengguna twitter lain dan berisi opini tentang berbagai topik yang terbatas biasanya dalam 140 karakter, sehingga twitter menjadi salah satu situs yang menyediakan kumpulan data opini dari masyarakat di seluruh dunia [28].

Selain itu, twitter mempunyai fitur *trending topic* yang menampilkan daftar topik terpopuler yang diperbarui setiap saat. Keuntungan lain dari twitter adalah semua postingan yang ada di twitter dapat dilihat oleh pengguna lain. Dengan kemudahan ini, pengguna twitter dapat saling berinteraksi tanpa harus menjadi pengikut. Keunggulan lainnya dari keterbukaan platform twitter yaitu dapat men-tweet dan menandai *(mention)* siapa saja tanpa harus membuat tweet di *fans page* atau akun pihak yang ditandai tersebut [18].

2.2.8 *Scraping*

Scraping adalah proses yang mengambil dokumen semi-terstruktur dari internet kemudian mengekstrak data spesifik dari halaman yang dapat digunakan untuk kepentingan lain. Manfaatnya adalah informasi yang diambil atau digunakan lebih terkonsentrasi sehingga lebih mudah dalam melakukan penelitian atau pencarian sesuatu. Proses *scraping* dari internet dapat dibagi menjadi dua langkah, yaitu menangkap sumber daya web dan kemudian mengekstrak informasi yang diinginkan dari data yang diambil. Secara khusus, program web *scraping* dimulai dengan meminta HTTP *(Hypertext Transfer – Transfer Protocol)* untuk mengembil sumber daya dari situs web target. Permintaan ini dapat diformat sebagai URL *(Uniform Resource Locator)* yang berisi permintaan GET atau HTTP yang berisi POST. Setelah permintaan berhasil diterima dan diproses oleh situs web target, sumber daya yang diminta akan diambil dari situs web dan kemudian dikirim kembali ke program web *scraping*. Sumber daya ini bisa dalam berbagai format, seperti halaman web yang dibangun dengan HTML *(HyperText Markup Language)*, XML *(Extensible Markup Language)* atau

JSON *(JavaScript Object Notation)*, atau data multimedia seperti gambar, audio, dan video [29].

2.2.9 *Text Preprocessing*

Preprocessing adalah proses mengubah data mentah menjadi data sesuai dengan prosedur mining yang akan dilakukan dan merupakan langkah terpenting dalam data mining [30]. Proses pembersihan pada data berupa teks yang dilakukan antara lain :

a. *Cleaning*

Pembersihan data dilakukan dengan cara menghilangkan terhadap kata dan karakter untuk mengurangi *noise*. Pada tahap ini, simbol, angka, dan tanda baca yang tidak diperlukan untuk analisis sentimen akan dihilangkan.

b. *Case Folding*

Case folding atau kapitalisasi merupakan tahap pengubahan huruf pada data tweets menjadi huruf kecil.

c. *Tokenization*

Tokenizing atau tokenisasi adalah proses memecah kalimat menjadi token dari teks tweet atau komponen komponen individual. Dengan dilakukan pemotongan akan memudahkan langkah selanjutnya karena pada proses-proses tersebut akan mencocokkan berdasarkan kata.

d. *Normalization*

Pada tahap ini berfungsi untuk mengubah kata yang tidak baku menjadi kata baku serta menghilangkan karakter yang berulang. Proses normalisasi data memiliki kumpulan kata yang tidak baku beserta pasangan kata yang baku yang disimpan dalam sebuah korpus.

e. *Stopword Removal*

Stopword removal merupakan proses menghilangkan kata-kata yang tidak relevan pada deskripsi dengan memeriksa apakah kata-kata hasil *parsing* dokumen termasuk di dalam daftar kata tidak penting *(stoplist)* atau tidak. Jika termasuk di dalam *stoplist* maka kata-kata

tersebut akan dihapus dari deskripsi dan kata-kata yang tersisa di dalam deskripsi dianggap sebagai kata-kata penting atau *keywords.*

f. Stemming

Stemming adalah proses mengubah sebuah kata ke dalam bentuk kata dasarnya dengan cara menghilangkan imbuhan pada kata dalam suatu dokumen. Tujuan dari *stemming* adalah untuk mengurangi jumlah kata yang unik membantu meningkatkan performa klasifikasi [31].

2.2.10 Kamus Lexicon Inset

Kamus Lexicon Inset merupakan sentimen lexicon berbahasa Indonesia yang dikategorikan dalam positif dan negatif yang dapat digunakan untuk menganalisis sentimen publik terhadap topik atau peristiwa tertentu. Kamus Inset disusun secara manual dengan memberi bobot dari setiap kata dan ditambahkan juga dengan kata sinonim beserta kata dasar [32].

2.2.11 *Term Frequency Inverse Document Frequency* (TF-IDF)

Metode TF-IDF merupakan metode penghitungan bobot dari masing masing kata yang paling umum digunakan dalam pencarian informasi. Metode ini juga dikenal dengan hasil yang efisien, mudah dan akurat. Metode *Term Frequency-Inverse Document Frequency* (TF-IDF) adalah cara pembobotan hubungan suatu kata *(term)* terhadap dokumen. TF-IDF ini adalah sebuah ukuran statistik yang digunakan untuk menilai seberapa penting sebuah kata di dalam sebuah dokumen atau dalam sekelompok kata.

Konsep dari *Term Frequency* (TF) menunjukkan bahwa semakin sering suatu term muncul pada sebuah dokumen maka semakin tinggi pula nilai bobot dari *term* itu sendiri. Proses *Inverse Document Frequency (IDF),* di sisi lain merupakan kebalikan dari proses TF. Semakin sering suatu *term* muncul dalam dalam proses IDF maka semakin kecil nilai bobot dari *term* itu sendiri [33].

Untuk menghitung nilai IDF, digunakan persamaan 2.1

$$\mathrm{Idf_t} = \log \frac{N}{df_t} \qquad (2.1)$$

Sementara untuk menghitung TF-IDF, digunakan persamaan 2.2:

$$Tf.idf_{t,d} = tf_{t,d} \times Idf_t \quad (2.2)$$

Dimana:

Idf_t = inversi frekuensi dokumen dari kata t

N = banyaknya dokumen

df_t = banyaknya dokumen yang mengandung kata t

$Tf.idf_{t,d}$ = nilai bobot kata t pada dokumen d

$tf_{t,d}$ = frekuensi kemunculan term t pada dokumen d

2.2.12 *Klasifikasi*

Teknik klasifikasi adalah salah satu dari teknik *data mining* yang termasuk dalam *supervised learning*. *Supervised learning* artinya proses klasifikasi menggunakan sebuah training dataset. Tujuannya adalah untuk memprediksi target dari beberapa atribut [34].

a. Naïve Bayes Classifier

Naïve Bayes merupakan metode klasifikasi berdasarkan teorema Bayes. Algoritma ini pertama kali dikemukakan oleh Thomas Bayes. Disebut Naïve karena kondisi antar atribut diasumsikan saling bebas. Metode klasifikasi ini tepat digunakan saat input dalam jumlah yang besar. Klasifikasi ini lebih sering dipakai karena kecepatan dan kesederhanaannya. Metode ini dikategorikan sebagai klasifikasi sederhana yang sering mencapai suatu performa yang setara dengan algoritma lain seperti *Decision Tree* dan *Neural Network Classifier*. Klasifikasi ini sangat memperhatikan tingginya akurasi serta kecepatan dalam memproses suatu data dalam jumlah besar.

Metode Naive bayes classifier memiliki dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses yaitu analisis terhadap sampel dokumen berupa pemilihan kosa kata, yaitu kata yang mungkin muncul dalam koleksi dokumen sampel hingga dapat serta menjadi representasi

dokumen. Selanjutnya adalah penentuan probabilitas tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan *term* yang muncul dalam dokumen yang diklasifikasi [15]. Rumus untuk Naive bayes classifier ada pada persamaan berikut:

$$P(c_i) = \frac{fd(c_i)}{|D|} \quad (2.3)$$

Keterangan:

$P(c_i)$ = Probabilitas c_i yang merupakan kategori kelas

$fd(c_i)$ = Jumlah dokumen c_i

$|D|$ = Jumlah data latih/dokumen

Setelah mendapatkan probabilitas dari setiap kelas, selanjutnya yaitu menghitung probabilitas setiap fitur pada setiap kelas dengan persamaan:

$$P(w_k|c_i) = \frac{P(w_k)xP(C_i|w_k)}{P(C_i)} \quad (2.4)$$

Keterangan:

$P(w_k|c_i)$ = *Posterior*, adalah kemunculan peluang pada kategori k tertentu ketika terdapat kemunculan kata i

$P(w_k)$ = *Prior*, adalah peluang kemunculan dokumen pada kategori k

$P(c_i|w_k)$ = *Likelihood* atau *conditional probability*, adalah peluang sebuah kata i masuk ke dalam kategori j

$P(c_i)$ = Probabilitas kemunculan suatu kata pada kelas

Kemudian langkah selanjutnya adalah perhitungan untuk menentukan probabilitas data uji dari masing-masing kelas berdasarkan dari proses *learning*. Nilai probabilitas yang terbesar akan dipilih :

$$Vmap = \underset{\{kelas\ 0, kelas\ 1\}}{argmax} \prod_{i=1}^{n} P(w_k|c_i)\ x\ \ P(c_i) \quad (2.5)$$

Keterangan:

$P(w_k|c_i)$ = *Posterior,* adalah kemunculan peluang pada kategori k tertentu ketika terdapat kemunculan kata i.

$P(c_i)$ = Probabilitas kemunculan suatu kata pada kelas.

b. Support Vector Machine

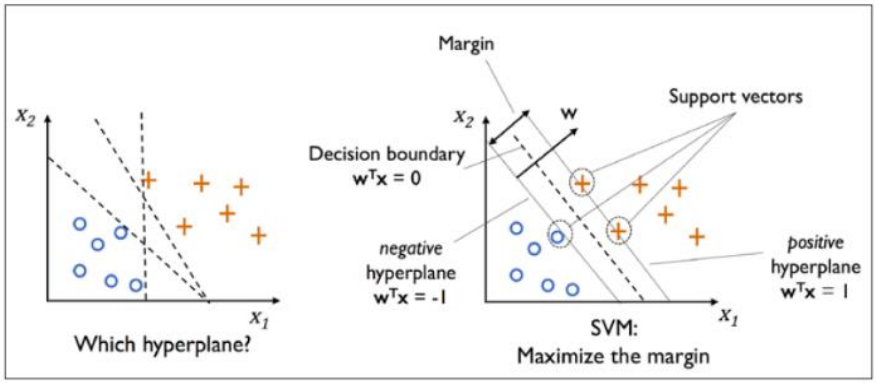
Support Vector Machine (SVM) adalah metode pembelajaran terbimbing yang menganalisis data dan mengenali pola yang digunakan untuk analisis klasifikasi dan regresi. Algoritma SVM pertama dibuat oleh Vladimir Vapnik dan turunan standar saat ini (margin lunak) diusulkan oleh Corinna Cortes dan Vapnik Vladimir [18]. SVM adalah algoritma pembelajaran mesin yang menerapkan fungsi *hyperplane* pada data untuk membentuk wilayah setiap kelas. *Hyperplane* sendiri merupakan sebuah fungsi yang digunakan sebagai pemisah antar kelas yang ada. [28]. Konsep inti dari *Support Vector Machine* adalah menarik garis yang optimal untuk mendapatkan *hyperplane* terbaik sebagai pemisah kedua kelas data dengan batas pemisah yang maksimum. Kelebihan dari algoritma *Support Vector Machine* adalah:

- Mempunyai kemampuan yang tinggi untuk menggeneralisasi
- Mampu menghasilkan model klasifikasi yang baik meskipun dilatih dengan himpunan data yang relatif kecil hanya dengan parameter yang sederhana. SVM juga memiliki konsep dan formula yang jelas dengan sedikit parameter yang harus diatur.

Sementara itu, kekurangan yang ada pada *Support Vector Machine* adalah sebagai berikut:

- Sulit diterapkan pada kumpulan data dengan jumlah sampel dan dimensi yang besar
- Secara umum dirumuskan hanya untuk menyelesaikan masalah klasifikasi dua kelas.

Dalam ruang kerja berdimensi tinggi, akan dicari *hyperplane* yang dapat memaksimalkan margin antara dua kelas. *Hyperplane* klasifikasi linear SVM dianotasikan sebagai berikut:



Gambar 2. 1 Hyperplane yang meisahkan dua kelas positif dan negatif

Untuk memperoleh garis *hyperplane* yang paling optimal dalam memisahkan data ke dua buah kelas tersebut, maka digunakan perhitungan *margin hyperplane* dan menemukan titik maksimal. Dalam memperoleh hyperplane pada SVM, dapat menggunakan persamaan (2.6)

$$(w.\ x_i) + b = 0 \tag{2.6}$$

Keterangan:

w = parameter *hyperplane* yang dicari

Di dalam data x_i , yang termasuk pada kelas -1 dapat dirumuskan seperti pada persamaan (2.7).

$$(w.\ x_i + b) \leq 1,\ yi = -1 \tag{2.7}$$

Sedangkan data x_i yang termasuk pada kelas +1 dapat dirumuskan seperti pada persamaan (2.8).

$$(w.\ x_i + b) \geq 1,\ y_i = 1 \tag{2.8}$$

Keterangan:

w= parameter *hyperplane* yang dicari

x= data input SVM

b= parameter *hyperplane* yang dicari (*bias*) [35].

2.2.13 Evaluasi

Evaluasi performansi dilakukan untuk mengetahui keakuratan dari pemodelan yang telah diterapkan pada data latih. Metode yang sering digunakan untuk evaluasi ada *confusion matrix*. *Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja metode klasifikasi. Tabel 2.2 menampilkan contoh *confusion matrix* pada dua kelas

Tabel 2.2 Confusion Matrix dua kelas

		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	*True Positive* (TP)	*False Positive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

Keterangan untuk tabel diatas dinyatakan sebagai berikut:

1. *True Positive* (TP), yaitu jumlah dokumen dari kelas positif yang benar dan diklasifikasikan sebagai kelas positif.
2. True Negative (TN), yaitu jumlah dokumen dari kelas negatif yang benar dan diklasifikasikan sebagai kelas negatif.
3. *False Positive* (FP), yaitu jumlah dokumen dari kelas negatif yang salah dan diklasifikasikan sebagai kelas positif.
4. *False Negative* (FN), yaitu jumlah dokumen dari kelas positif yang salah dan diklasifikasikan sebagai kelas negatif.

Kemudian menghitung *accuracy precision, recall,* dan *f1-score*.

a. Akurasi

Akurasi merupakan persentase dari keseluruhan sentimen yang benar dikenali. Perhitungan akurasi dilakukan dengan cara membagi jumlah data sentimen yang benar dengan total data uji. Untuk

menghitung nilai akurasinya dilakukan dengan menggunakan persamaan (2.9)

$$Akurasi = \frac{true\ positive}{jumlah\ data\ tes}\ x\ 100\% \quad (2.9)$$

b. Precision

Precision menggambarkan keakuratan model yang diminta dengan hasil prediksi yang diberikan. Perhitungan *precision* dilakukan dengan cara membagi jumlah data benar yang bernilai positif dibagi dengan jumlah data benar yang bernilai positif dan data salah yang bernilai positif. Nilai dari data salah bernilai positif diambil dari jumlah nilai selain *true positive* kolom yang sesuai tiap kelasnya. Untuk menghitung nilai *precision* dapat dilakukan dengan menggunakan persamaan (2.10)

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (2.10)$$

c. Recall

Recall menggambarkan keberhasilan model dalam menemukan kembali informasi. Perhitugan *recall* dilakukan dengan membagi data benar bernilai positif dengan hasil penjumlahan dari data benar yang bernilai positif dan data salah yang bernilai negatif. Nilai dari data salah yang bernilai negatif diambil dari jumlah nilai selain *true positive* baris yang sesuai tiap kelasnya. Perhitungan *recall* dapat menggunakan persamaan (2.11)

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (2.11)$$

d. F-1 Score

F1-Score merupakan parameter tunggal ukuran keberhasilan retrieval yang menggabungkan *recall* dan *precision*. Nilai *F1-score* adalah hasil perkalian *precision* dan *recall* dibagi dengan hasil penjumlahan *precision* dan *recall* kemudian dikalikan dua. Perhitungan *f1-score* menggunakan persamaan (2.12). [36]

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \quad (2.12)$$

2.2.14 K-Fold Cross Validation

K-Fold Cross Validation adalah salah satu teknik untuk mengevaluasi atau memvalidasi keakuratan sebuah model yang dibangun berdasarkan dataset tertentu. Pembuatan model biasanya bertujuan untuk membuat prediksi maupun klasifikasi terhadap suatu data baru yang mungkin tidak muncul dalam dataset. Salah satu metode *cross validation* yang paling populer adalah *K-Fold cross validation*. K-*fold* bekerja dengan melipat data sebanyak K dan mengulangi (meng-iterasi) nya sebanyak K dan mengulangi eksperimenya sebanyak K juga [37].

2.2.15 Python

Python adalah pemrograman interpretatif multifungsi dengan menggunakan perancangan yang berfokus pada tingkat keterbacaan kode. Keuntungan dari python yaitu kode dari python lebih mudah dipahami oleh programmer pemula. Python di kembangkan oleh Guido van Rossum dari Amsterdam, Belanda dan pertama kali menjadi domain publik pada tahun 1991. Bahasa pemrograman ini di gunakan untuk membuat perangkat lunak bahkan beberapa perusahaan menggunakan bahasa pemrograman python untuk membuat perangkat lunak komersial. Fitur-fitur yang dimiliki oleh python yaitu memiliki *library* yang luas sehingga menyediakan modul-modul yang bisa digunakan sesuai kebutuhan, memiliki bahasa yang mudah di pelajari, memiliki *layout* kode sumber yang memudahkan untuk melakukan pemeriksaan kembali pada penulisan kode sumber, memiliki sistem pengolahan memori otomatis, mudah untuk di kembangkan dan menciptakan modul-modul baru [38].