BAB II
TINJAUAN PUSTAKA

Pada bab ini penulis akan menjelaskan tentang ringkasan tertulis dari beberapa jurnal, artikel, dan buku sebagai bahan penelitian lain yang diarahkan untuk menyusun konsep yang berkaitan dengan peneliti dari penjelasan studi-studi sebelumnya dan dasar-dasar teori yang digunakan.

2.1. Penelitian Terdahulu

Pada bagian ini terdapat peninjauan terhadap penelitian sebelumnya dalam area yang sama yang telah dilakukan oleh peneliti lain sebelumnya. Selama melakukan penelitian, penulis telah mempelajari berbagai hal dari literatur lain yang dapat dimasukkan dalam laporan. Penelitian seharusnya memperlihatkan cara pendekatan yang digunakan oleh peneliti sebelumnya dalam menanggapi permasalahan yang sama melalui studi literatur. Semua bahan bacaan yang dipakai dalam bagian ini perlu diacu.

Studi ini menggunakan jurnal dan berbagai referensi yang relevan sebagai sumber informasi untuk keperluan penelitian. Referensi yang digunakan mencakup sumber-sumber yang terkait dengan pengkajian sentimen melalui web maupun aplikasi Android. Penulis mengutip sumber-sumber lain yang juga merujuk pada beberapa tugas akhir yang membahas metode Support Vector Machine.

2.2. Jurnal Penelitian

Tabel 2. 1 Matriks literatur riview dan perbandingan penelitian

No.	Pengarang	Judul	Tahun	Kekurangan	Kelebihan	Kesimpulan dan Saran
1	Alberi Meidharma Fadli Hulu, Kemas Muslim Lhaksmana	Analisis sentimen politik pada twitter menggunakan metode	2019	Jumlah data yang di gunakanhanya sebesar 1000 data *tweet.*	Pemilihan metode Support Vector Machine dipilih karena memiliki kemampuan	Sentimen yang paling banyak muncul oleh penggunna twitter mengenai pilpres 2019

Tabel 2.1 Lanjutan

No.	Pengarang	Judul	Tahun	Kekurangan	Kelebihan	Kesimpulan dan Saran
		Support Vector Machine.			generalisasi dalam mengklasifikasi suatu pattern dengan akurasiyang cukup tinggi.	adalah sentimen yang bernilai psitif, dalam pengembangan kedepannya, dapat dilakukan penambahan jumlah tweet
2	Putri Yuniasari, Febri Maspiyanti	Analisis sentimen data tweet menggunakan Metode Support Vector Machine (Studi Kasus: PemindahanIbukota Baru Republik Indonesia)	2021	Sentimen dengan kata negasi belum dapat ditentukan polaritasnya dengan optimal.	Menggunakan metode SVM maka hasil hipotesa awal sesuai hasil akhir yakni SVM lebih baik dari metode sebelumnya.	Bobot tweet didapat dari rankscore yang dihasilkan oleh *Support Vector Machine* (SVM).Saran yang didapatkan dari penelitian yang dilakukan oleh penulis adalah untuk penelitian mengembangkan aplikasi ini menggunakan algoritma klasifikasi yang lainnya.
3	Ilham taufik	Analisis sentiment terhadap tokoh public menggunakan algoritma SVM	2018	Tidak ada pemilihan daftar stop-list	Hasil menunjukkan bahwa kernel linier memiliki tingkat presisi paling baik yaitu 80%	Analisa sentiment pada tweet bekerja dengan baik dilihat dari hasilakurasi ,penelitihanya menggunakan 2kelas yaitu positif dan negatif selanjutnya dapat diteliti bagaimana

Tabel 2.1 Lanjutan

No.	Pengarang	Judul	Tahun	Kekurangan	Kelebihan	Kesimpulan dan Saran
						implementasi metode SVM untuk mengklasifikasikan teks multiclass
4	M Indra Halim ArsyaDwi Akbari, DKK	Analisis sentimen menggunakan metode Learing Vector	2017	Tidak adanya teori penangana n nagasi untuk lebih menganalisa keambiguan dalam tweet	Dari hasil yang diperoleh Penggunaan metode LVQ baik dalam menganalisis sentiment	Hasil paling baik didapatkan dengan menggunakan learing rate 0.2 pada 100 data dengan hasil 74.8%.menambahkan frekuensi kata dalam kamus sehingga kosa kata dari kata-kata tidak baku bisa tertangani dengan baik
5	Gery Nugroho, Danang Triantoro Murdiansyah, Kemas M Lhaksmana	Analisis sentimen pemilihan presiden amerika 2020 di twitter menggunakan naïve bayes dan support vector machine	2021		Penelitian mencoba membandingkan algoritma SVM dan naïve bayes	Proses analisis sentiment pada pemilihan presiden amerika 2020lebih baik menggunakan Support Vector Machine, mencoba mengunakan metode klasifikasi lainnya yang belum di gunakan pada penelitian ini.

2.3. Analisis Sentimen

Analisis sentimen atau bisa bahasa lainnya opinion mining merupakan sebuah

teknik untuk menentukan sentimen atau pendapat seseorang yang di analisis dalam bentuk teks dan dapat diklasifikasikan dengan kategori teks tersebut positif, negatif, atau netral. (Arief 2019)

Analisis sentimen menjadi salah satu topik penelitian dalam kategori Natural Language Processing sejak awal tahun 2000. Tujuan dari penelitian ini adalah untuk membuat sebuah metode otomatis yang diimplementasikan menjadi tools yang dimanfaatkan untuk mengekstrasi informasi subjectif , seperti pendapat atau opini dari berbagai topik, produk, layanan, atau aktifitaslainnya menjadi dalam bentuk sebuah data.(Liu, B. 2012)

2.4. Twitter

Twitter, salah satu media sosial paling populer di dunia, telah menjadi sumber data yang berharga untuk analisis sentimen dalam berbagai konteks, termasuk politik. Karakteristik unik Twitter, seperti batasan karakter dalam tweet, memberikan keunggulan dalam identifikasi sentimen dalam pesan-pesan singkat (Thelwall, Buckley, & Paltoglou, 2012). Selain itu, Twitter menyediakan akses API yang memungkinkan peneliti untuk mengumpulkan data tweet dalam jumlah besar dan secara otomatis (Yanuarsyah et al., 2018).

Pengguna Twitter aktif dalam berbicara tentang politik dan pemilihan umum, menjadikan platform ini relevan untuk menganalisis sentimen politik (Suryadinata & Budiarto, 2019). Data dari Twitter dapat memberikan wawasan berharga tentang pandangan masyarakat terhadap kandidat, partai politik, dan isu-isu politik selama kampanye pemilihan (Pratama et al., 2021).

Penelitian sebelumnya telah berhasil menggunakan data Twitter untuk analisis sentimen dalam berbagai konteks politik, yang menegaskan validitas dan relevansi platform ini (Yanuarsyah et al., 2018; Pratama et al., 2021).

2.5. Twitter API

API atau kepanjangan dari Application Programming Interfacemerupakan antarmuka yang digunakan untuk memungkinkan satu aplikasi berkomunikasi atau bertukar data dengan yang lain. Twitter merupakan salah satu sosial media yang menyediakan API untuk keperluan penelitian atau bertukar informasi data untuk

pengguna umum.(Widodo, 2021).

Twitter API terbentuk pada November 2016 , dimana pada saat itu hanya dapat menampilkan data tempat, negara, profile pengguna twitter, hingga negara pengguna twtter. Namun sekarang pada website restem Developer platform milik Twitter yaitu https://developer.twitter.com/en, sekarang twitter API sudah mencapai versi 2.0.

2.6. Natural Language Processing (NLP)

Bidang kecerdasan buatan yang berkaitan dengan bahasa alami manusia seperti bahasa Inggris atau bahasa Indonesia disebut Natural Language Processing atau NLP. Bagian dari ilmu komputer ini berfokus pada interaksi antara komputer dan bahasa manusia. Tujuan utama dari ilmu NLP adalah untuk menghasilkan kemampuan bagi mesin untuk memahami dan memperoleh pemahaman tentang makna bahasa manusia, sehingga memungkinkan mesin untuk memberikan respon yang tepat dan sesuai. (Alamanda, R. S. 2016).

Secara kasar NLP dapat mencari siapa melakukan apa kepada siapa, kapan, di mana, bagaimana dan mengapa. Hal tersebut karena NLP dapat mengekstrak kata bebas lebih jauh dari suatu teks bebas.(Wangsanegara, N. K. 2015).

Berdasarkan (David, P. L. 2010) NLP memiliki 3 aspek pada teori pemahaman.

1. Syntax

Aspek ini menjelaskan bahasa dari bentuk kata. Syntax biasa diten-tukan dari sebuah grammer, hal ini karena bahasa alami atau bahasa sehari – hari jauh lebih sulit daripada bahasa formal untuk diimplementasikan kedalam kecerdasan buatan dan program komputer.

2. Semantics

Pada *semantics* menjelaskan kalimat dalam satu bahasa. Dalam penerapan NLP untuk aplikasi tertentu , aspek ini diperlukan untuk menganalisis makna teks yang lebih sederhana.

3. Pragmatics

Pada aspek ini berkaitan dengan komunikasi dan interpretasi bahasasecara umum. Hal Ini melibatkan penggunaan bahasa yang bermakna dalam berbagai

situasi.

2.7. Text mining

Text mining merupakan suatu metode yang digunakan untuk mengungkap dan mengekstrak informasi dari sejumlah dokumen dengan memanfaatkan perangkat analisis yang termasuk dalam konsep data mining. Salah satu tahapan dalam metode ini adalah kategorisasi.(Ma'arif, A. A. 2015).

Text mining dapat memberikan solusi dari memproses, mengelompokkan, dan menganalisis teks yang tidak terstuktur dalam jumlah data yang besar. Text mining memiliki keterkaitan dengan bidang ilmu yang lain seperti Data mining, Information Retrieval, , Machine Learning, NLP, dan Visualization.

Mirip dengan data mining, text mining juga merupakan metode algoritma yang secara terstruktur memproses data teks dengan beberapa langkah. Secara keseluruhan, dalam text mining terdapat tiga tahapan utama yang meliputi preprocessing teks, seleksi fitur, dan analisis teks. (Priyanto, A. 2018).

2.8. Preprocessing

Preprocessing atau praproses merupakan proses untuk mempersiapkan data atau teks yang tidak terstruktur menjadi terstruktur agar dapat digunakanproses selanjutnya.(Joang, I. K. 2017).

pada preprocessing memiliki beberapa tahapan yaitu case folding, cleansing, tokenizing, stopword, dan stemming dengan tujuan memperbaiki data mentah menjadi data yang dapat diolah dalam sistem (Mujilahwati, S. 2016), berikut penjelasan dari beberapa tahapan tersebut :

1. Case folding

   Case folding merupakan tahapan awal dalam preprocessing. Case folding memiliki tujuan untuk mengubah setiap bentuk kata yang sama yaitu dengan melakukan perubahan kata menjadi lower case atau huruf kecil. case folding hanya berlaku untuk huruf alfabet.(Athira Luqyana, W. C. 2018)

2. Cleansing

   Pada tahapan ini merupakan proses untuk membersihkan kata-kata yang tidak diperlukan seperti menghilangkan delimiter tanda koma "," , tanda titik "." ,

dan tanda baca lainnya. selain menghilangkan delimiter proses ini menghilangkan karakter HTML, kata kunci, ikon emosi, hashtag (#), username (@username), url (http://website.com), dan email (nama@website.com), tujuannya adalah untuk mengurangi noise yang terdapat pada teks.(Buntoro, B. C. 2017).

3. Tokenizing

Tokenize memiliki fungsi untuk memecah dokumen menjadi beberapa kumpulan kata. Tokenizing dapat dilakukan dengan menghilangkan tanda baca dan memisahkannya berdasarkan spasi.(Wahyuni, R. T. 2015).

4. Stopword

Istilah stopword adalah kata yang tidak relevan dengan topik utama yang berasal dari basis data meskipun sering ditemukan dalam dokumen. Beberapa kata atau kelompok kata dalam bahasa Indonesia yang sering digunakan sebagai stopwords antara lain: yang, juga, dari, dia, kami, kamu, aku, saya, ini, itu, atau, dan, ke, tak, tidak, di, dll. (Susilowati, E. K. 2015).

5. Stemming

Stemming merupakan proses pre-processing untuk mencari katadasar dari hasil stopword removal. Stemming dilakukan dengan cara pendekatan kamus atau dengan pendekatan aturan. (Athira Luqyana, W. C. 2018).

Selain 5 tahapan sebelumnya terdapat tahapan tambahan yaitu seleksi kata slang. Kata slang adalah kata bahasa yang tidak resmi atau tidak memenuhi standar kamus Indonesia (KBBI) dan tidak baku yang sifatnya musiman, kata slang digunakan oleh kaum remaja atau kelompok sosial tertentu untuk komunikasi intern.(Riyaddulloh, R. 2021).

Kata slang biasanya dalam bentuk singkatan atau istilah gaul yang muncul di masyarakat, kata slang sering ditemukan hampir disemua media sosial.(Khomsah, S. 2017). Contohnya dari kata slang seperti kata “b aja” yang memiliki arti “biasa aja” dan masih banyak lagi. Berikut beberapa dataset yang dapat digunakan dalam menerapkan seleksi kata slang.

1. Dataset bahasa alay milik Okky Ibrohim

Dataset yang dibuat oleh (Ibrohim, M. O. 2018) memiliki 238 kata slang yang dapat digunakan untuk melakukan seleksi kata slang.

2. Dataset bahasa alay milik Salsabila

Dataset dari penelitian (Salsabila, N. A. 2018) memiliki 15.007 kata slang yang dapat digunakan untuk melakukan seleksi kata slang.

2.9. Seleksi fitur

Seleksi fitur merupakan tahapan penting karena dapat mempengaruhi tingkat akurasi klasifikasi. Dalam penerapan seleksi fitur tidak boleh sembarangan karena jika dataset berisi sejumlah fitur, maka dimensi ruang akan menjadi besar sehinga menyebabkan tingkat akurasi menjadi rendah. Seleksi fitur dapat mempengaruhi beberapa aspek seperti pola klasifikasi, akurasi klasifikasi, waktu yang diperlukan untuk pembelajaran fungsi klasifikasi. (Arifin, 2016).

Dalam seleksi fitur terbagi dua kategori yaitu seleksi fitur supervised dan seleksi fitur unsupervised. Contoh dari seleksi fitur supervised adalah chi square, information gain dan mutual information. Sedangkan untuk contoh seleksi fitur unsupervised adalah term strength, term contribution, entropy based ranking dan document frequency. perbedaan dari kategori tersebut adalah keberadaan informasi awal tentang kategori dari suatu dokumen.(Tsani et al., 2020).

2.9.1. TF-IDF

TF-IDF adalah sebuah metode yang digunakan untuk menunjukkan seberapa sering suatu kata muncul dalam sebuah dokumen. Metode TF-IDF adalah gabungan dari dua metode, yakni TF (Term Frequency) dan IDF (Inverse Document Frequency). Term Frequency adalah pengukuran seberapa sering sebuah kata muncul dalam sebuah dokumen, sedangkan Inverse Document Frequency (IDF) adalah sebuah teknik yang digunakan untuk mengurangi bobot kata-kata yang sering muncul dalam dokumen, karena kata-kata tersebut dianggap umum dan tidak penting.(Rizkia, S. S. 2019).

$$IDF\ (w) = \log(\frac{n}{DF(w)}) \qquad (2.1)$$

Rumus 2. 1 Fungsi Idf

$$TF - IDF(w,d) = TF(w,d) \times IDF(w) \qquad (2.2)$$

Rumus 2. 2 Fungsi TF-Idf

Keterangan

TF-IDF (w,d) = bobot kata dalam dokumen

W = kata

D = dokumen

TF (w,d) = frekuensi kemunculan kata (w) dalam dokumen (d)

IDF (w) = inverse DF (Document Frequency) dari sebuah kata (w)

N = banyaknya data atau dokumen

DF = banyaknya kata (w) dalam dokumen (d)

2.10.Klasifikasi

Klasifikasi adalah sebuah teknik dalam machine learning yang digunakan untuk mengategorikan atau mengelompokkan objek berdasarkan karakteristik tertentu, mirip dengan cara manusia dalam membedakan satu benda dengan lainnya.(Ahmad, A. 2017). Metode klasifikasi memiliki beberapa algoritma yang sering digunakan dalam penelitian machine learning yaituDecision/classification trees, Bayesian classifiers/NaïveBayes classifiers, Support vector machines (SVM), Algoritma Genetika, Rough sets, k-nearest neighbor, Metode Rule Based, Memory based reasoning, ,Analisa Statistik, dan Neural networks.(Leidiyana, H. 2013).

Dalam melakukan proses klasifikasi terdapat 4 komponen fundamental yaitu (Gorunescu, F. 2011):

1. Kelas

Merupakan variabel kategori yang terikat dalam model dan mem-presentasikan label pada suatu objek. Contoh customer loyalty, jenis gempa atau badai, jenis rasi bintang, dll.

2. Prediktor

Merupakan variabel independen dari model yang mewakilkan karak-ter (atribut) dari data yang akan di klasifikasi.

3. Training dataset

Merupakan kumpulan data yang berisi nilai dari 2 komponen sebelumnya yaitu data kelas dan data prediktor. Training bertujuan untuk melatih model mengenali kelas berdasarkan prediktor yang tersedia.

4. Testing dataset

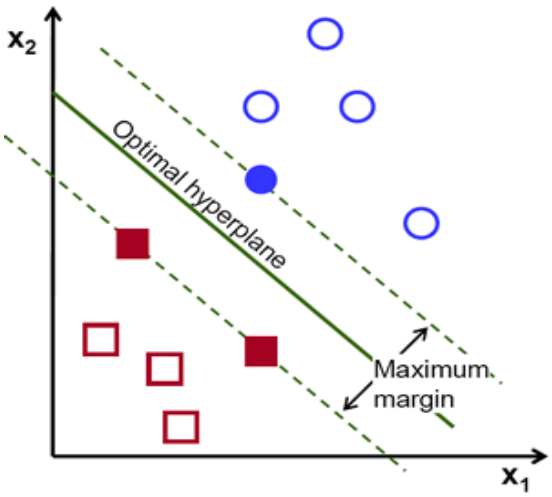
Merupakan data baru yang akan diklasifikasi oleh model yang telah dibangun melalui training data agar mendapatkan akurasi yang dapat dievaluasi.

2.11.Support vector machine

Support Vector Machine (SVM) adalah suatu teknik klasifikasi untuk melakukan prediksi. Support Vector Machine termasuk kategori klasifikasi supervised learning, dimana dalam implementasi diperlukan tahap pelatihan menggunakan sequential training SVM dan disusul tahap pengujian. (Rofiqohet, U. P. 2017)

Algoritma SVM diperkenalkan oleh Vapnik, Boser, dan Guyon pada tahun 1992 dan memiliki perkembangan yang signifikan saat itu. Pada Support vector machine, konsep dasarnya adalah menggunakan garis pembatas (hyperlane) untuk memisahkan dua kelas. Garis pembatas yang baik adalah garis yang memiliki jarak terbesar ke titik data pelatihan terdekat dari setiap kelas. Margin yang semakin besar umumnya akan menghasilkan error generalisasi yang lebih rendah. Margin adalah jarak antara titik vektor dalam kelas tertentu terhadap hiperplane.(Sari, B. W. 2019).

Konsep dasar pada support vector machine dapat di ilustrasikan dalam gambar 2. 1.



Gambar 2. 1 Ilustrasi algoritma Support Vector Machine

Pada algoritma svm memiliki persamaan sebagai berikut. (Rofiqoh, U. P. 2017)

$$f(x) = w, x + b \qquad (2.3)$$

Rumus 2. 3 Persamaan Algoritma SVM

atau

$$f(x) = \sum_{t=1}^{m} a_i y_i \, K(x, x_i) + b \qquad (2.4)$$

Rumus 2. 4 Persamaan Algoritma SVM

Keterangan :

1. w : parameter hyperplane yang dicari (garis yang tegak lurus antaragaris hyperplane dan titik support vector)
2. x : titik data masukan Support Vector Machine
3. ai : nilai bobot setiap titik data
4. $K(x, xi)$: fungsi kernel
5. b : parameter hyperplane yang dicari (nilai bias)

2.12.Confusion matrix

Confusion matrix merupakan metode yang berfungsi untuk menganalisis seberapa baik hasil klasifikasi dalam mengenali kumpulan data dari kelas yang berbeda.(Romadoni, F. U. 2020). Analisis yang dilakukan dalam metodeconfusion matrix meliputi perhitungan akurasi, recall, precision, dan error rate.(Rizki, S. S. 2019). Berikut contoh confusion matrix dengan 2 kelas yaitu positif dan negatif.

Tabel 2. 2 Confusion Matriks 2 kelas

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FP
	Negatif	FN	TN

1. Recall

Recall adalah suatu metode penghitungan yang digunakan untuk mengukur sejauh mana kasus yang benar dapat diidentifikasi dengan benar. rumus yang digunakan untuk menghitung nilai recall, yaitu sebagai berikut.

$$Recall = \frac{TP}{TP+FN} \qquad (2.5)$$

Rumus 2. 5 Persamaan menentukan recall

2. Precission

Precision adalah rasio perbandingan antara prediksi true positif (TP)dengan semua hasil prediksi berniilai positif (FP). Persamaan untuk melakukan perhitungan precision sebagai berikut.

$$Precission\ \frac{TP}{TP+FP} \qquad (2.6)$$

Rumus 2. 6 Persamaan menentukan precision

3. Error Rate

Error rate merupakan perhitungan rasio antara salah satu hasil prediksi dengan seluruh data. Persamaan untuk melakukan perhitungan *errorrate* sebagai berikut.

$$Error\ rate = \frac{TP+FN}{TN+FP+FN+TP} \qquad (2.7)$$

Rumus 2. 7 Persamaan menentukan error rate

4. Akurasi

Akurasi adalah perhitungan rasio prediksi *true* positif (TP) dan *true*negatif (TN) dengan seluruh data. Persamaan untuk melakukan akurasi sebagai berikut.

$$Accuracy = \frac{TP+TN}{TN+FP+FN+TP} \qquad (2.8)$$

Rumus 2. 8 Persamaan menentukan akurasi

Keterangan :

1. TP = Data kelas positif dengan hasil prediksi positif
2. FP = Data kelas negatif dengan hasil prediksi positif
3. FN = Data kelas positif dengan hasil prediksi negatif
4. TN = Data kelas negatif dengan hasil prediksi negatif