BAB II. LANDASAN TEORI

2.1 Studi Literatur

Beberapa penelitian terdahulu yang menjadi rujukan dalam penelitian ini adalah sebagai berikut :

Tabel 2. 1 Studi Literatur

No.	Judul	Penulis	Metode	Data yang Digunakan	Hasil Penelitian
1.	Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naive Bayes Classifier Berbasis Konversi Ikon Emosi.	Fransiska Vinasari dan Arief Wibowo (2019)	Naive Bayes Classifier	Data yang digunakan berasal dari media sosial twitter dengan jumlah data 900 tweet, 300 tweet positif, 300 tweet netral dan 300 tweet negatif. Pengumpulan data tweet pada twitter menggunakan RStudio dengan memasukan settingan twitter API dan mengetikan kata kunci JD.id.	Hasil penelitian menunjukkan bahwa metode Naïve Bayes tanpa penambahan fitur mampu mengklasifikasi sentimen dengan nilai akurasi sebesar 96,44%, sementara jika ditambahkan fitur pembobotan tf-idf disertai konversi ikon emosi mampu meningkatkan nilai akurasi menjadi 98%.
2.	Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine	Valentino Kevin Sitanayah Que, dkk (2020)	Support Vector Machine berbasis Particle Swarm	Data tweet dengan metode scraping menggunakan Octoparse. Total 1.852 data tweet dari 01 Januari 2019 hingga 15	Analisis sentimen positif menggunakan SVM adalah sebesar 62% dan sentimen negatif sebesar 38%, sedangkan pada SVM-PSO, opini positif sebesar 53% dan negatif 47%. Hasil penelitian

	Berbasis Particle Swarm Optimization.		Optimizati on.	Oktober 2019 yang dibagi menjadi data testing 1.130 tweet dan training 722 tweet.	menggunakan 10 k-fold CV menghasilkan akurasi pada SVM sebesar 95,46% dan AUC 0,979 (excellent classification), sedangkan pada SVM-PSO sebesar 96,04% dan AUC 0,993 (excellent classification). Hasil menunjukkan bahwa penggunaan data training dan testing dapat dilakukan dan terbukti bahwa SVM-PSO lebih baik daripada SVM biasa, meskipun menggunakan nilai parameter default.
3.	Analisis Sentimen Twitter Menggunakan Metode Naive Bayes dengan Relevance Frequency Fitur Selection (Studi Kasus: Opini Masyarakat Mengenai Kebijakan New Normal).	Kresentia Verena Septiana Toy, dkk. (2021)	Naive Bayes dengan Relevancy Frequency Fitur Seletion.	Dataset yang digunakan adalah 300 data opini masyarakat yang didapat dari sosial media twitter.	Hasil dari pengujian sebanyak 5 pengujian menggunakan klasifikasi Naive Bayes, diperoleh rata-rata akurasi sebesar 62,6%, sementara hasil pengujian akurasi klasifikasi dengan penambahan RFFS diperoleh ratarata akurasi sebesar 65,3%.
4.	Analisis Sentimen Kebijakan	Dhaifa Farah Zhafira,	Naive Bayes dengan	Data yang digunakan didapat dari kolom	5 proses utama dalam penelitian ini yang meliputi pelabelan manual, text

	Kampus Merdeka Menggunakan Naive Bayes dan pembobotan TF-IDF Berdasarkan Komentar pada Youtube.	dkk. (2021)	Pembobotan TF-IDF.	komentar konten yang diunggah pada youtube dengan 900 data latih dan 100 data uji.	preprocessing, pembobotan TF-IDF, validasi data menggunakan k-fold cross validation, dan klasifikasi. Hasil akurasi terbaik sebesar 97% yang didapat dengan menggunakan 900 data latih, 100 data uji, menerapkan pembobotan TF-IDF, dan 10-fold cross validation. Rata-rata akurasi yang didapat dari 10 iterasi pada k-fold cross validation yaitu sebesar 91.8% dengan nilai precision, recall, f-measure sebesar 90.35%, 93.6%, 91.95%. Berdasarkan hasil tersebut, Naive Bayes Classifier cukup baik sebagai alternatif untuk analisis sentimen.
5.	Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Wordcloud dengan Metode Naive Bayes dan K-Nearest Neighbor	Josua Josen A. Limbong, dkk (2021)	Naive Bayes dan K-Nearest Neighbor	penelitian ini menggunakan data ulasan sebanyak 500 ulasan yang didapat dari website google play store.	Hasilnya setelah menggunakan metode Naive Bayes memperoleh hasil nilai accuracy 0,914, precision 0,915, recall 0,914 dan F1 score 0,916. Sedangkan metode KNN memperoleh nilai accuracy 0,928, precision 0,929, recall 0,928, dan F1 score 0,926. Hal ini membuktikan bahwa dalam penelitian ini kinerja metode KNN lebih baik. Kemudian

					berdasarkan hasil word cloud yang diperoleh didapatkan informasi kata dengan sentimen positif yang paling sering diulas oleh pelanggan diantaranya terkait kata: gratis, bagus, suka, murah, mudah, dan cepat. Sedangkan informasi sentimen negatif yang diperoleh seperti kata : kecewa, jelek, mahal, bohong, ribet, dan perbaiki.

2.2 Dasar Teori

2.2.1 Text Mining

Text mining dapat didefinisikan secara luas sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi. Text mining adalah sebuah proses ekstraksi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar teks, seperti dokumen Word,PDF,kutipan teks, sedangkan input untuk data mining adalah data yang terstruktur (Feldman & Sanger, 2007).

2.2.2 Analisis Sentimen

Analisis Sentimen atau *Sentiment Analysis* (SA) merupakan proses memahami dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat atau teks yang berupa opini. Tujuan dilakukan *sentiment analysis* untuk melihat pandangan atau pendapat teks yang berkaitan terhadap sebuah masalah atau objek, apakah cenderung berpandangan positif atau negatif. *Sentiment analysis*

terdiri dari pemrosesan bahasa alami, 8 analisis teks dan komputasi linguistik untuk mengidentifikasi sentimen dari suatu dokumen (Vinodhini & Chandrasekaran, 2015).

2.2.3 E-Commerce

E-commerce adalah suatu proses transaksi yang dilakukan oleh pembeli dan penjual dalam membeli dan menjual berbagai produk secara elektronik dari perusahaan ke perusahaan lain dengan menggunakan komputer sebagai perantara transaksi bisnis yang dilakukan (Loudon, 1998).

E-commerce adalah aktivitas belanja online dengan menggunakan jaringan internet serta cara transaksinya melalui transfer uang secara digital (Prawiro, 2021).

2.2.4 Shopee

Shopee adalah aplikasi *online shop* atau *marketplace* (platform perdagangan elektronik). Dengan menggunakan shopee, kita bisa lebih mudah berbelanja, menjelajah, dan menjual produk serta jasa apa saja dan dimana saja. Shopee dapat membantu para penjual lebih mudah menawarkan barang dagangan mereka dan membantu pembeli dalam melakukan transaksi serta berinteraksi langsung dengan para penjual melalui fitur *live chat*nya (Akbar, 2021).

2.2.5 Text Preprocessing

Text preprocessing adalah suatu proses untuk menyeleksi data *text* agar menjadi lebih terstruktur lagi dengan melalui serangkaian tahapan yang meliputi tahapan *case folding*, *tokenizing*, *filtering* dan *stemming*. *Text preprocessing* merupakan salah satu implementasi dari *text mining*. *Text mining* sendiri adalah suatu kegiatan menambang data, dimana data yang biasanya diambil berupa *text* yang bersumber dari dokumen-dokumen yang memiliki *goals* untuk mencari kata kunci yang mewakili dari sekumpulan

dokumen tersebut sehingga nantinya dapat dilakukan analisa hubungan antara dokumen-dokumen tersebut (Tineges, 2021).

2.2.6 TF-IDF (Term Frequency – Inverse Document Frequency)

Term Frequency — Inverse Document Frequency atau TF — IDF adalah suatu metode algoritma yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat. Metode ini akan menghitung nilai *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Secara sederhana, metode TF-IDF digunakan untuk mengetahui berapa sering suatu kata muncul di dalam dokumen.

$$TF = jumlah\ frekuensi\ kata\ terpilih\ /\ jumlah\ kata \quad (1)$$

$$IDF = \log(jumlah\ dokumen\ /\ jumlah\ frekuensi\ kata\ terpilih) \quad (2)$$

TF menggunakan perbandingan antara frekuensi sebuah *term* dengan nilai maksimum dari keseluruhan atau kumpulan frekuensi *term* yang ada pada suatu dokumen. Metode IDF merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan (Delta, 2019).

2.2.7 Naive Bayes

Naive Bayes merupakan suatu metode yang berguna untuk melakukan pengklasifikasian dengan aturan-aturan Bayes yang ada menggunakan perhitungan peluang. Pada proses klasifikasi dilakukan pengkategorian menggunakan nilai probabilitas maksimal. Metode Naive Bayes sering digunakan para peneliti karena dirasa efektif untuk memperoleh ketepatan hasil dengan akurasi yang tinggi (Kusumadewi, 2009).

Pemilihan Metode *Naive Bayes Classifier* dikarenakan beberapa kelebihan yang alogritma tersebut miliki, kelebihan tersebut antara lain :

- Algoritma *Naive Bayes* bekerja sangat cepat dan dengan mudah memprediksi kelas dari kumpulan data pengujian.
- Pengklasifikasian *Naive bayes* bekerja lebih baik daripada model lain dengan lebih sedikit data latih.
- Bisa digunakan untuk data kuantitatif maupun kualitatif.
- Jika ada nilai yang hilang, maka bisa diabaikan dalam perhitungan.
- Perhitungannya cepat dan efisien.
- Tidak memerlukan data *training* yang banyak.
- Bisa digunakan untuk klasifikasi masalah biner ataupun *multiclass*.

Metode *Naive Bayes* dapat digunakan untuk beberapa aplikasi umum dengan klasifikasi. Seperti :

- Prediksi *real-time*. Karena kemudahan implementasi dan komputasi yang cepat, metode ini bisa digunakan untuk melakukan prediksi secara *real-time*.
- Prediksi multi-kelas. Metode *Naive Bayes* dapat digunakan untuk memprediksi probabilitas posterior dari beberapa kelas variabel target.
- Klasifikasi teks. Karena fitur prediksi multi-kelas, *Naive Bayes* sangat cocok untuk klasifikasi teks yang digunakan untuk memecahkan masalah seperti penyaringan spam dan analisis sentimen.
- Sistem rekomendasi. Dengan adanya algoritma penyaringan kolaboratif, *Naive Bayes* membuat sistem rekomendasi yang dapat digunakan untuk menyaring informasi tidak terlihat dan untuk memprediksi apakah pengguna menyukai hasil yang diberikan atau tidak.

Tahapan proses algoritma *Naive Bayes* adalah sebagai berikut :

a) Membaca data *training*.
b) Hitung jumlah *class*.
c) Hitung jumlah kasus yang sama dengan *class* yang sama.
d) Kalikan semua nilai hasil sesuai dengan data X yang dicari *class* nya.
e) Bandingkan hasil tiap kelas.

Persamaan teorema *Bayes* sebagai berikut (Limbong, Sembiring, & Hartomo, 2021) :

$$P(H|X) = \frac{P(H) * P(X|H)}{P(X)} \quad (3)$$

Keterangan:

X = Data dengan *class* yang belum diketahui

H = Hipotesis data X yang merupakan suatu *class* yang lebih spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (posteriori *probability*)

P(H) = Probabilitas hipotesis H (prior *probability*)

P(X|H) = Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) = Probabilitas X.

2.2.8 Confusion matrix

Confusion matrix digunakan sebagai teknik evaluasi pada penelitian ini. *Confusion matrix* merupakan metode evaluasi yang terdiri dari hasil prediksi oleh sistem dengan hasil yang aktual atau sebenarnya. Tabel 1 merupakan tabel *confusion matrix* (Zhafira, Rahayudi, & Indriati, 2021).

Tabel 2. 2 Confusion Matrix

		Prediksi	
		Negatif	Positif
Aktual	Negatif	*True Negative* (TN)	*False Positive* (FP)
	Positif	*False Negative* (FN)	*True Positive* (TP)

Keterangan :

a) *True Negative* (TN) merupakan jumlah kalimat negatif yang dideteksi sebagai kalimat negatif.

b) *False Negative* (FN) merupakan jumlah kalimat positif yang dideteksi sebagai kalimat negatif.
c) *False Positive* (FP) merupakan jumlah kalimat negatif yang dideteksi sebagai kalimat positif.
d) *True Positive* (TP) merupakan jumlah kalimat positif yang dideteksi sebagai kalimat positif.

Dari tabel diatas, terdapat beberapa metrik yang dihitung untuk mengevaluasi algoritma yaitu *accuracy*, *precision*, *recall*, dan *F1 score*. Dalam *confusion matrix*, terdapat empat keterangan (Islamy, Adikara, & Indriati, 2022):

- *True Positive* (TP) merupakan jumlah kalimat positif yang dideteksi sebagai kalimat positif.
- *False Positive* (FP) adalah merupakan jumlah kalimat negatif yang dideteksi sebagai kalimat positif.
- *True Negative* (TN) yang merupakan merupakan jumlah kalimat negatif yang dideteksi sebagai kalimat negatif.
- *False Negative* (FN) merupakan jumlah kalimat positif yang dideteksi sebagai kalimat negatif.

Accuracy adalah jumlah rasio dari nilai prediksi benar dengan jumlah nilai semua data. Akurasi dapat dihitung dengan menggunakan persamaan berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision adalah perbandingan antara jumlah kalimat positif yang benar dideteksi sebagai kalimat positif dengan jumlah kalimat positif yang dideteksi. Persamaan rumus untuk perhitungan *precision* adalah

$$precision = \frac{TP}{TP + FP} \quad (5)$$

Recall merupakan perbandingan antara jumlah kalimat positif yang benar dideteksi sebagai kalimat positif dengan jumlah data aktual yang berupa kalimat positif. Rumus persamaan untuk perhitungan *recall* adalah

$$recall = \frac{TP}{TP + FN} \tag{6}$$

F1 Score adalah metrik pengukuran yang menggabungkan antara *precision* dan *recall*. *F1 Score* menghitung kinerja keseluruhan klasifikasi. Rumus persamaan untuk perhitungan *F1 score* adalah

$$F1\ Score = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

2.2.9 Php

PHP adalah skrip bersifat server-side yang ditambahkan ke halaman HTML. Skrip ini akan membuat suatu aplikasi dapat di integrasikan ke dalam HTML sehingga suatu halaman web tidak lagi bersifat statis, namun menjadi dinamis. Sifat server side berarti pengerjaan kode program dilakukan di server, baru kemudian hasilnya di kirimkan ke browser (Kustiyaningsih, 2011).