

## **BAB II**

### **TINJAUAN PUSTAKA**

#### **2.1 Kajian Pustaka**

Alven Safik Ritonga dan Endah Supeni Purwaningsih tahun 2018 melakukan penelitian dengan mengklasifikasikan hasil pengelasan untuk membantu proses pemantauan kualitas. Peneliti menggunakan data hasil pengujian dari PT. MorteK Suraindo dengan 3 variabel yaitu Uji Kekuatan Tarik (T), Uji Struktur Makro (M), Uji Kekerasan *Rockwell*. *Dataset* yang digunakan kemudian dikategorikan dengan 3 kelas yaitu kelas 1, kelas 2 dan kelas 3. Data yang digunakan sebanyak 100 baris data dengan pembagian data latih dan data tes sebesar 75:25. Berdasarkan hasil pengujian didapatkan tingkat akurasi sebesar 98% dengan menggunakan kernel fungsi quadratic [9].

Go-Pay merupakan salah satu layanan yang diberikan oleh Go-Jek yang bergerak dalam bidang penyedia layanan uang elektronik. Seiring dengan pemakaian pada masyarakat kemunculan Go-Pay menimbulkan respon dari masyarakat. Respon masyarakat tidak semuanya berupa respon positif terhadap layanan Go-Pay terdapat juga respon negatif. Penelitian yang dilakukan oleh Rachmad Mahendrajaya, Ghulam Asrofi Buntoro dan Moh. Bhanu Setyawan pada 2019 menganalisis sentimen masyarakat terhadap pengguna Go-Pay. Penelitian ini menggunakan data dari Twitter sebanyak 1210 baris data. *Dataset* tersebut dianalisis dengan metode pelabelan *lexicon based* kemudian diklasifikasikan dengan metode Support Vector Machine. Dari hasil pengujian dengan Confusion Matrix didapatkan hasil akurasi sebesar 89,17% dengan jumlah ulasan positif sebanyak 923 dan 287 ulasan negatif [7].

Penelitian yang dilakukan oleh Elly Indrayuni dengan topik sentiment analisis terhadap review produk kosmetik telah dilakukan pada tahun 2019. Produk kosmetik saat ini menjadi kebutuhan utama bagi para perempuan. Banyaknya produk kosmetik akan memunculkan berbagai review baik positif ataupun negatif. Pengguna merasa kesulitan jika harus membaca komentar satu

per satu sehingga dibutuhkan pengelompokan komentar secara otomatis. Peneliti mengambil data dari *website* yang berbahasa Indonesia yaitu <https://femaledaily.com/>. Dari *website* tersebut diambil sebanyak 200 komentar yang terdiri dari 100 komentar positif dan 100 komentar negatif. Pemilihan komentar positif dan negatif tersebut dilakukan peneliti secara manual. Pengolahan data digunakan metode Naive Bayes yang dioptimalkan dengan menambah proses N-Gram pada tahap *preprocessing*. Dari hasil pengujian yang dilakukan didapatkan tingkat akurasi sebesar 90.50% dan nilai AUC sebesar 0.715 [5].

Penelitian yang dilakukan oleh Sulastris dan Eri Zuliarso dengan topik mengklasifikasikan dokumen putusan kasasi pengadilan tentang merk. Isi teks dokumen yang teratur akan mempengaruhi kerja sistem klasifikasi, hal ini disebabkan oleh sistem klasifikasi menggunakan isi teks sebagai dasar pertimbangan, Data yang digunakan dalam penelitian ini berupa dokumen berbahasa Indonesia yaitu dokumen putusan Pengadilan tentang merk yang diambil dari <https://putusan.mahkamahagung.go.id/pengadilan/mahkamah-agung/direktori/perdata-khusus/Merk>. Data yang digunakan sebanyak 50 buah, yang terdiri dari 40 buah data latih dan 10 data uji. Pada data latih terdapat 2 buah label yaitu tolak dan Kabul dengan pembagian masing-masing label sebesar 50%. Data diolah pada proses pre-processing untuk mendapatkan data yang matang kemudian masing-masing data diklasifikasikan menggunakan metode Naïve Bayes. Berdasarkan hasil pengujian didapatkan hasil akurasi sebesar 80% [10].

Rian Tinages, Agung Triayudi dan Ira Diana Sholihati pada tahun 2020 melakukan penelitian terhadap sentimen analisis pelanggan Indihome. Penelitian ini dilakukan menggunakan data dari Twitter, diambil dari *tweet* yang menyebutkan *username* @IndiHome sejumlah 1000 baris data kemudian dilakukan tahap *preprocessing* dan didapatkan *dataset* matang sebanyak 1400 baris. Tahap analisis dilakukan menggunakan metode Support Vector Machine dengan menggunakan fitur ekstraksi TF-IDF. Berdasarkan model yang dibangun dengan menggunakan 1120 baris data latih kemudian dilakukan pengujian

dengan Confusion Matrix. Hasil Confusion Matrix didapatkan tingkat akurasi sebesar 87% dengan 367 *tweet* memiliki polaritas positif dan 83 *tweet* memiliki polaritas negatif [8].

Penelitian yang dilakukan oleh Siti Azza Amira, Satria Utama , dan Muhammad Hanif Fahmi untuk menganalisa ulasan hotel pada salah satu layanan *booking* hotel *online*. Seiring dengan berkembangnya internet yang semakin pesat maka kemudahan mendapatkan informasi akan semakin meningkat. Informasi yang dapat dicari adalah ulasan mengenai bagaimana kualitas hotel yang ditawarkan secara online. Salah satu faktor penentu calon pelanggan dalam menentukan tempat menginapnya adalah dengan melihat bagaimana ulasan tempat tersebut. Tripadvisor merupakan *website* yang menyediakan layanan pemesanan hotel dan menampilkan ulasan dari pengguna sebelumnya. Peneliti menggunakan data berupa teks berbahasa Indonesia dari ulasan di *website* tripadvisor. Untuk menentukan polaritas dari masing-masing data digunakan metode Support Vector Machine (SVM) dan pembobotan kata menggunakan Term Frequency Inverse Document Frequency (TF-IDF). TF-IDF mampu memberikan bobot pada masing-masing kata sehingga dapat dijadikan bahan perhitungan pada metode klasifikasi SVM. Hasil penelitian didapatkan tingkat akurasi 88% , presisi sebesar 98% dan *recall* sebesar 89% [11].

Penelitian yang dilakukan oleh Helena Nurramdhani Irmanda dan Ria Astriratma dengan topik klasifikasi jenis pantun. Pantun adalah sebuah puisi lama dari ilmu sastra Indonesia. Pantun dapat digunakan untuk berkomunikasi dengan menyelipkan nasihat, pertanyaan, jawaban maupun kritik di dalamnya. Menurut isi teksnya pantun terdiri dari beberapa kategori yaitu pantun anak-anak, pantun muda dan pantun tua. Dalam penelitian ini dilakukan klasifikasi terhadap 470 buah pantun kedalam tiga kategori diatas. Pembuatan model dilakukan menggunakan 80% dari seluruh data dan 20% data digunakan sebagai pengujian model. Pembuatan model dilakukan dengan menggunakan metode SVM dengan pembobotan katanya menggunakan TF-IDF. Data uji akan diklasifikasikan menggunakan model yang sudah dibuat kemudian akan dibandingkan kembali hasil klasifikasinya dengan kelas sebenarnya. Dari

pengujian tersebut maka akan dapat diamati tingkat keakurasian dari model yang telah dibuat. Berdasarkan hasil pengujian peneliti mendapatkan tingkat akurasi sebesar 91.91% [12].

Penelitian yang dilakukan oleh Taofik Krisdiyanto dan Erry Maricha Oki Nurharyanto dengan topik sentimen analisis masyarakat terhadap penerapan PPKM. Pada awal tahun 2021 pemerintah mengeluarkan sebuah kebijakan untuk menanggulangi penyebaran COVID-19 dengan dibatasinya kegiatan masyarakat yang dikenal dengan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM). Kebijakan tersebut menimbulkan berbagai respon masyarakat di media sosial, salah satunya adalah Twitter. Penelitian ini bertujuan untuk melihat bagaimana sentimen masyarakat terhadap penerapan kebijakan PPKM di Indonesia. Data yang digunakan berupa teks dari tagar “ppkm” di Twitter sejumlah 1000 baris data. Data tersebut dilakukan tahap pra-processing untuk mendapatkan data yang matang, kemudian diklasifikasikan kedalam 2 klasifikasi yaitu positif dan negatif. Untuk mengklasifikasikan data tersebut peneliti menggunakan metode Naïve Bayes. Hasil penelitian didapatkan 99% data tweet dari masyarakat termasuk kedalam polaritas positif dan sisanya polaritas negatif. Berdasarkan seluruh data kata terbanyak yang paling sering muncul adalah kata darurat [6].

Penelitian yang dilakukan oleh Ulvah dan Aishiyah Saputri Laswi pada tahun 2022 menganalisis data kelangkaan minyak goreng dengan metode apriori. Algoritma apriori adalah salah satu algoritma yang banyak dipakai untuk mencari hubungan antara item satu dengan item lainnya. Algoritma ini adalah salah satu algoritma dalam *data mining* yang sering disebut dengan analisis keranjang belanja. Algoritma ini dapat digunakan juga sebagai pertimbangan dalam mengatur tata letak *display* produk sehingga pelanggan akan mudah dalam menemukan barang. Berdasarkan penelitian yang dilakukan didapatkan hasil 13 jenis barang yang sering dibeli secara bersamaan tiap terjadi transaksi berdasarkan 210 jenis barang yang tersedia pada toko [13].

Tabel 2. 1 Tabel Penelitian Terdahulu

No	Judul	Tahun	Masalah	Metode	Hasil	Perbedaan
1	Penerapan Metode Support Vector Machine (Svm) Dalam Klasifikasi Kualitas Pengelasan Smaw (Shield Metal Arc Welding)	2018	Mengklasifikasi kualitas hasil las SMAW pada PT. MorteK Suraindo.	Support Vector Machine	Berdasarkan hasil pengujian didapatkan tingkat akurasi sebesar 98% dengan menggunakan kernel fungsi kuadrat	Penelitian ini tidak menggunakan data berupa teks tetapi menggunakan data kategorikal
2	Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based Dan Support Vector Machine	2019	Menganalisis sentimen respon masyarakat terhadap pelayanan Go-Pay di Indonesia	<i>Lexicon Based</i> dan Support Vector Machine	Dari hasil pengujian dengan Confusion Matrix didapatkan hasil akurasi sebesar 89,17% dengan jumlah ulasan positif sebanyak 923 dan 287 ulasan negatif.	Pembobotan dokumen pada penelitian ini menggunakan metode Term Document Metrix.
3	Klasifikasi Text Mining Review Produk Kosmetik	2019	Klasifikasi ulasan produk kosmetik	Naïve Bayes	Berdasarkan hasil pengujian menggunakan	Penelitian dalam tahap preprocessing tidak terdapat langkah konversi <i>three or</i>

No	Judul	Tahun	Masalah	Metode	Hasil	Perbedaan
	Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes				Confusion Matrix didapatkan hasil akurasi sebesar 90.50%.	<i>more</i> dan konversi kata slang.
4	Klasifikasi Dokumen Dengan Metode Naive Bayes Terhadap Putusan Kasasi Pengadilan Tentang Merk	2020	Klasifikasi dokumen keputusan kasasi pengadilan tentang merk	Naive Bayes	Penelitian ini mendapatkan hasil akurasi sebesar 80% dengan metode pengujian <i>confusin matrix</i> .	Pembobotan dokumen pada penelitian ini menggunakan metode Term Document Metrix.
5	Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (Svm)	2020	Menganalisis sentimen respon masyarakat terhadap pelayanan Indihome di Indonesia	Support Vector Machine	Dari 1400 data matang yang dianalisis didapatkan akurasi 87% dengan <i>tweet</i> dengan polaritas positif sebanyak 367 dan 87 <i>tweet</i> berpolaritas negatif.	Penelitian ini belum dijelaskan mengenai tahap pelabelan data.
6	Penerapan Metode Support Vector Machine Untuk Analisis Sentimen	2020	Menganalisis sentimen ulasan pelanggan hotel yang diambil dari	Support Vector Machine	Pengujian yang dilakukan menggunakan Confusion	Penelitian ini belum dijelaskan mengenai tahap pelabelan data.

No	Judul	Tahun	Masalah	Metode	Hasil	Perbedaan
	Pada Review Pelanggan Hotel		<i>website</i> . Hasil analisis dapat digunakan sebagai bahan pertimbangan calon pelanggan untuk menggunakan hotel.		Matrix didapatkan hasil akurasi 88%. Dari 223 baris data didapatkan 177 data memiliki polaritas positif dan 44 data memiliki polaritas negatif	
7	Klasifikasi Jenis Pantun Dengan Metode Support Vector Machine (SVM)	2020	Pengklasifikasian pantun ke dalam 3 jenis pantun yaitu pantun anak-anak, pantun muda dan pantun orang tua	Support Vector Machine	Berdasarkan hasil pengujian menggunakan Confusion Matrix, didapatkan hasil akurasi sebesar 81.91%. hasil tersebut didapatkan dari 470 baris data dengan proporsi 376 baris data latih dan 94 baris data uji.	Pelabelan dalam penelitian ini tidak berbasis <i>lexicon based</i> .

No	Judul	Tahun	Masalah	Metode	Hasil	Perbedaan
8	Analisis Sentimen Opini Masyarakat Indonesia Terhadap Kebijakan PPKM Pada Media Sosial Twitter Menggunakan Naïve Bayes Clasifiers	2021	Menganalisis sentimen pendapat masyarakat mengenai kebijakan pemerintah dala menekan angka penularan Covid-19 dengan cara menerapkan PPKM.	Naïve Bayes	Berdasarkan 1000 data yang digunakan dalam proses analisis didapatkan hasil 99% data memiliki polaritas positif sedangkan 1% memiliki polaritas negatif. Kata terbanyak yang muncul dalam <i>dataset</i> adalah kata “darurat”.	Dalam penelitian ini tidak dijelaskan mengenai tahap palabelan <i>dataset</i> serta tahap pembobotan. Metode yang diguanakand alampenelitian ini adalah Naive Bayes.
9	Implementasi Algoritma Apriori Dalam Keterkaitan Data Pada Kelangkaan Minyak Goreng	2022	Kelangkaan minyak goreng menyebabkan pembeli susah mendapatkan minyak goreng sehingga perlu dilakukan analisis untuk menentukan keputusan penjualan minyak goreng	Apriori	Dari 210 produk terdapat 13 produk yang biasanya dibeli barsamaan dengan minyak goreng	Penelitian ini tidak menggunakan metode Naïve Bayes maupun SVM.



No	Judul	Tahun	Masalah	Metode	Hasil	Perbedaan
			berdasarkan barang lain yang dibeli.			

Berdasarkan penelitian - penelitian yang sudah dilakukan di atas metode Naïve Bayes dan SVM digunakan sebagai metode pengklasifikasian dengan sebagian besar *dataset* yang digunakan adalah data teks. Dari hasil pengujian yang dilakukan kedua metode tersebut mampu menghasilkan tingkat akurasi yang tinggi. Sehingga metode tersebut cocok untuk dijadikan sebagai metode penelitian untuk menganalisis sentimen masyarakat terhadap mahalnya harga minyak goreng. Penelitian yang dilakukan oleh Elly Indrayuni dan penelitian yang dilakukan oleh Rachmad Mahendrajaya akan dijadikan sebagai referensi utama dalam penelitian yang akan dilakukan.

## **2.2 Landasan Teori**

### **2.2.1 Twitter**

Berbagai media sosial digunakan oleh masyarakat Indonesia saat ini. Twitter merupakan salah satu pilihan dari beberapa media sosial yang legal di Indonesia. Jack Dorsey pada Maret tahun 2006 mendirikan Twitter dan pertama kali diluncurkan pada bulan Juli 2006. Saat itu Twitter berhasil berada di sepuluh besar media sosial yang paling sering digunakan sejak diluncurkan. Twitter memberikan sebuah fitur yaitu pengguna dapat membuat cuitan / *tweet*. Dalam menuliskan cuitan, pengguna dapat menggunakan bahasa mereka sehari-hari sehingga jika diamati sebagian besar cuitan menggunakan bahasa yang tidak baku. Bahasa yang tidak baku tersebut membuat karakteristik data yang didapatkan dari Twitter bersifat tidak terstruktur [14].

### **2.2.2 Analisis Sentimen**

Pengolahan Bahasa alami terdiri dari beberapa bagian salah satunya adalah analisis sentimen. Analisis sentiment berguna dalam pengolahan Bahasa alami, *text mining* serta komputasi linguistik dimana hal tersebut digunakan untuk menentukan opini pada sebuah topik. Opini yang dihasilkan tersebut dapat mengindikasikan penilaian, alasan dan kondisi kecerendungan. Analisis sentimen banyak digunakan untuk menganalisis opini masyarakat terhadap suatu barang atau jasa.

Analisis sentimen menghasilkan dua atau lebih polaritas, umumnya adalah polaritas positif dan polaritas negatif. Polaritas tersebut dapat digunakan sebagai bahan pertimbangan untuk dijadikan sebuah keputusan. Naïve Bayes dan SVM merupakan metode yang dapat digunakan dalam analisis sentimen [11].

### 2.2.3 Naïve Bayes

Naïve Bayes adalah metode klasifikasi yang diambil dari teorema Bayes. Klasifikasi menggunakan metode Naïve Bayes banyak digunakan karena dapat digunakan untuk data yang besar serta kecepatan dalam menganalisis data. Metode Naïve Bayes adalah klasifikasi yang sederhana untuk mengklasifikasikan teks. Model ini adalah bentuk sederhana dari Bayesian Network. Metode ini bekerja dengan cara melakukan analisis terhadap data latih yang sudah terdapat label di masing-masing baris data. Algoritma hasil pelatihan atau model dapat digunakan untuk mengklasifikasikan data baru berdasarkan polaritas yang ada.

Bentuk dasar teorema Bayes dapat diamati dalam persamaan 2.1. *Posterior probability* adalah peluang terjadinya H jika terdapat sebuah *Evidence A*. *Likelihood* adalah probabilitas A terhadap hipotesis H. *Prior* adalah probabilitas hipotesis tanpa dipengaruhi oleh hal apapun. Nilai *Evidence* dalam praktiknya selalu bernilai konstan sehingga dapat diabaikan. Perhitungan hanya terfokus terhadap perkalian antara nilai *prior* dan nilai *likelihood* ketika *evidence* diabaikan [15].

$$P(H|A) = \frac{P(A|H)P(H)}{P(A)} ; P(A) \neq 0 \quad (2.1)$$

Keterangan:

$P(H|A)$  : *posterior probability*

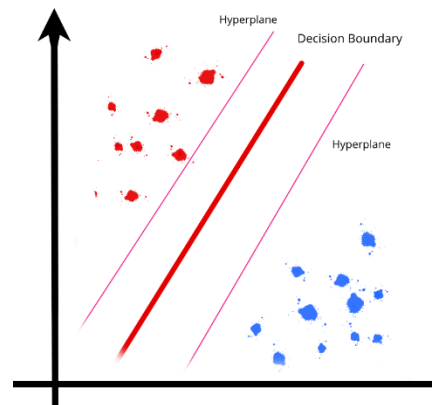
$P(H)$  : *prior probability*

$P(A|H)$  : *likelihood*

$P(A)$  : *evidence*

### 2.2.4 Support Vector Machine (SVM)

Support Vector Machine pada dasarnya bekerja dengan *linear classifier* atau bisa disebut dengan klasifikasi biner. Sebagai bagian dari supervised learning Support Vector Machine akan bekerja dengan data latih yang sudah memiliki label kelas yang jelas. Hasil dari pelatihan Support Vector Machine akan mampu memprediksi data baru akan masuk kedalam kelas tertentu. Untuk menentukan hasil prediksi metode Support Vector Machine menggunakan *decision boundary* atau batas keputusan, ilustrasi mengenai *decision boundary* dapat diamati dalam gambar 3.1.



Gambar 2. 1 Gambar *Decision Boundary*

Metode Support Vector Machine juga mampu melakukan klasifikasi non linear yang dilakukan dengan menggunakan kernel. Klasifikasi dengan menggunakan fungsi kernel digunakan untuk membuat ruang fitur berdimensi lebih tinggi dengan menjadikan data non linear menjadi terpisah secara linear. Fungsi kernel yang dapat digunakan dalam klasifikasi menggunakan Support Vector Machine diantaranya: [16] [17]

#### 1. Kernel Polynomial

$$k(\hat{x}_i, \hat{x}_j) = (\hat{x}_i \hat{x}_j + 1)^p \quad (2.2)$$

Keterangan :

$\hat{x}_i$  = data latih

$\hat{x}_i$  = data uji  
 $p$  = derajat polynomial

## 2. Kernel Gaussian RBF

$$k(\hat{x}_i, \hat{x}_j) = \exp\left(-\frac{\|\hat{x}_i - \hat{x}_j\|^2}{2\sigma^2}\right) \quad (2.3)$$

Keterangan :

$\hat{x}_i$  = data latih  
 $\hat{x}_j$  = data uji  
 $\sigma$  = parameter pada RBF

## 3. Kernel Sigmoid

$$k(\hat{x}_i, \hat{x}_j) = \tanh(\alpha \hat{x}_i \hat{x}_j + \beta) \quad (2.4)$$

Keterangan:

$\hat{x}_i$  = data latih  
 $\hat{x}_j$  = data uji  
 $\alpha$  = parameter sigmoid  
 $\beta$  = koefisien

Derajat polynomial pada hasil klasifikasi digunakan untuk mengontrol fleksibilitas *decision boundary*. Semakin tinggi nilai derajat polynomial dapat membuat penurunan akurasi karena model yang terlalu fokus pada data latih. Nilai gamma pada kernel sigmoid dapat menurunkan nilai akurasi jika digunakan nilai yang terlalu tinggi, tetapi akurasi dipengaruhi juga oleh jumlah fitur, semakin banyak jumlah fitur maka digunakan nilai gamma yang kecil [17].

### 2.2.5 Confusion Matrix

Tahap pengujian model yang telah dibangun dilakukan menggunakan pendekatan Confusion Matrix. Confusion Matrix adalah metode untuk mengukur hasil sebuah model dengan pengukuran akurasi, *precision*, *recall* dan *F1-Score*. Akurasi adalah menghitung seberapa

persen semua kelas yang diprediksi benar negatif dan positif dari keseluruhan data. *Precision* adalah menghitung berapa persen sebuah kelas diprediksi dengan benar dibandingkan dengan keseluruhan prediksi pada kelas tersebut. *Recall* adalah membagi antara jumlah klasifikasi yang benar dengan total keseluruhan aktual. *F1-Score* adalah perbandingan rata-rata *precision* dan *recall* yang di bobotkan. Secara umum bentuk Confusion Matrix dapat diamati pada tabel 2.2 [18].

Tabel 2. 2 Tabel Confusion Matrix

		NILAI SEBENARNYA	
		Positif	Negatif
PREDIKSI	Positif	<i>TruePs</i>	<i>FalsePs</i>
	Negatif	<i>FalseNg</i>	<i>TrueNg</i>

Berdasarkan tabel diatas maka nilai akurasi dapat diamati dalam persamaan 2.5.

$$accuracy = \frac{TruePs + TrueNg}{\sum data} \quad (2.5)$$

Untuk mendapatkan nilai *precision* dibutuhkan perhitungan *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Perhitungan tersebut dilakukan terhadap masing-masing kelas. Untuk mencari nilai *precision* dapat diamati dalam persamaan 2.6.

$$precision = \frac{TruePs}{TruePs + FalsePs} \quad (2.6)$$

Untuk menghitung *Recall* dapat diamati dalam persamaan 2.7:

$$recall = \frac{TruePs}{TruePs + FalseNg} \quad (2.7)$$

. Untuk menghitung nilai *F1-Score* dapat di amati pada persamaan 2.8.

$$F1\ Score = \frac{2 * precision * recall}{precision + recall} \quad (2.8)$$