

BAB II

TINJAUAN PUSTAKA

2.1. Analisis Sentimen

Analisis sentimen atau pertambangan pendapat adalah studi komputasi pendapat, perasaan dan emosi diekspresikan dalam teks. Segmen *review* di iPhone untuk memperkenalkan masalah (nomor dikaitkan dengan setiap kalimat untuk referensi mudah):

“(1) Saya membeli iPhone beberapa hari yang lalu. (2) Itu seperti telepon bagus. (3) Layar sentuh benar-benar keren. (4) Kualitas suara jelas juga. (5) Meskipun baterai yang tidak lama, itu ok bagi saya. (6) Namun, ibu saya marah dengan saya karena saya tidak mengatakan padanya sebelum saya membelinya. (7) Dia juga berpikir telepon itu terlalu mahal, dan ingin saya untuk mengembalikannya ke toko. ...”

Pertanyaannya adalah: apa yang kita ingin tambang atau ekstrak dari ulasan ini? Hal pertama yang kita mungkin melihat adalah bahwa ada beberapa pendapat dalam ulasan ini. Kalimat (2), (3) dan (4) mengungkapkan pendapat positif, sedangkan kalimat (5), (6) dan (7) mengungkapkan pendapat negatif atau emosi. Pendapat semua memiliki beberapa target atau benda yang pendapat diekspresikan. Pendapat dalam kalimat (2) adalah pada iPhone secara keseluruhan, dan pendapat dalam kalimat (3), (4) dan (5) berada di “layar sentuh”, “kualitas suara” dan “hidup baterai” dari iPhone masing-masing. Pendapat dalam kalimat (7) adalah pada harga iPhone, tapi pendapat atau emosi dalam kalimat (6) adalah pada “aku”, bukan

iPhone. Ini adalah poin penting. Sebuah aplikasi, pengguna dapat tertarik opini yang objektif tentang target atau benda tertentu, namun tidak pada semua (misalnya, mungkin pada “aku”). Memerhatikan sumber-sumber atau pemegang pendapat. Sumber atau pemegang pendapat dalam kalimat (2), (3), (4) dan (5) adalah penulis review (“I”), tetapi dalam kalimat (6) dan (7) adalah “ibu saya”. Dengan contoh ini dalam pikiran, kita sekarang secara resmi menentukan analisis sentimen atau masalah pertambangan pendapat (Liu, 2010).

2.2. Twitter

Twitter adalah situs *web* yang dioperasikan dan dimiliki oleh Twitter Inc. Twitter merupakan aplikasi media sosial yang sering dipakai oleh penggunanya dalam menyampaikan opininya (Jtik et al., 2022). Inti dari twitter adalah *tweets*. Pada aplikasi twitter, pengguna atau user diberikan fitur untuk bisa mengirimkan tweet dengan batas 140 karakter. Tweet tidak boleh lebih dari 140 karakter. Oleh karena itu, keterbatasan ini membuat banyak pengguna yang sering melakukan penyingkatan saat membuat tweet agar apa yang ia kemukakan bisa tertuangkan secara keseluruhan (Of et al., 2021) .

2.3. Text Data Mining

Text data mining atau disingkat dengan TDM memiliki nama lain yaitu *Knowledge Discovery in Text*. *Text data mining* ini dikembangkan secara khusus agar bisa mengekstraksi informasi dari susunana kata-kata yang tidak terstruktur dari dokumen. Dokumen yang berisi kata-kata akan menjadi sumber data yang

memiliki tujuan untuk mencari kata-kata yang dapat menjadi perwakilan dari isi dokumen tersebut. Kata-kata yang didapat akan dianalisis untuk melihat hubungan antara dokumen (Prasetyo, 2014).

Text data mining akan berkonsentrasi pada metodologi atau juga teknik yang bermuat kembali pada ruang lingkup sistem informasi. Text mining merupakan sebuah proses yang membuat pengguna berinteraksi dengan sekumpulan dokumen-dokumen dengan menggunakan beberapa teknik yang ada dalam analisis (Efendi & Mustakim, 2017). Tahapan yang paling utama dalam melakukan metode *text mining* adalah dengan menemukan kata-kata yang menjadi perwakilan dari setiap isi dokumen. Selanjutnya akan mencari dengan melakukan analisis terhadap hubungan satu dokumen dengan dokumen yang lain menggunakan beberapa teknik yaitu analisis klasifikasi, kelompok atau asosiasi. Tahapan awal yang digunakan dalam *text mining* adalah *text preprocessing*. Text data mining juga bisa diartikan sebagai metode yang bisa digunakan sebagai pengambil keputusan, analisis dan juga mengembangkan teknik pemrosesan data berupa text dalam jumlah yang cukup besar (Tuffery, 2011).

2.4. Text Preprocessing

Text preprocessing merupakan tahapan yang penting dalam mengelola data sebelum masuk ke dalam proses klasifikasi. Adapun lima tahap text preprocessing dalam penelitian ini adalah sebagai berikut (Nuansa, 2017) :

1. *Case Folding* merupakan proses pada text preprocessing yang mana pada proses ini dokumen atau text yang terdiri dari susunan huruf akan dirubah

menjadi huruf kecil secara keseluruhan supaya tidak terjadi ketidaksamaan di dalam dokumen.

2. *Cleansing* merupakan teknik yang akan membuat dokumen atau text menjadi bersih dari karakter yang tidak diperlukan seperti *hashtag*, *username*, *url*, *emoticons* dan *HTML*. Pembersihan ini berguna agar noise bisa berkurang sehingga proses klasifikasi bisa berjalan dengan baik.
3. *Tokenizing* merupakan proses dimana dokumen yang tersusun dari kata-kata akan dipecah menjadi kata-kata yang akan memutus urutan string.
4. *Stopwords*, yakni membuang kosakata yang tidak memiliki arti khusus atau kata-kata unik yang bisa menjadi perwakilan dari dokumen (Dragut et al., 2009). Contoh kosakata tersebut adalah dan, ini, itu, adalah dan lain sebagainya.
5. *Stemming* merupakan teknik yang berguna untuk mengembalikan kata ke kata dasar karena menghilangkan kata awalan, sisipan, kombinasi dan juga akhiran. Contohnya seperti kata mengambil maka dalam teknik *stemming* akan berubah menjadi ambil.

2.5. Machine Learning

Machine learning merupakan suatu bidang atau ruang lingkup yang mana di dalamnya mempelajari berbagai pengembangan dari bermacam-macam algoritma komputer untuk mengubah sesuatu data menjadi sebuah informasi yang berguna. Teknik-teknik serta algoritma dalam data mining berguna untuk menjadi model pembelajaran yang dapat memprediksi sesuatu yang akan terjadi selanjutnya dari

data yang sudah ada. Dalam ruang lingkup machine learning ada dua tipe yang biasa digunakan yaitu *unsupervised learning* dan *supervised learning*. Ada beberapa contoh algoritma pada tipe *supervised learning* yaitu Naïve Bayes, *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Neural Network*, *Decision Trees*, *Linear Regression* dan lain sebagainya (Faizal, 2016).

2.6. Klasifikasi

Klasifikasi menjadi salah satu cabang dari data mining. Klasifikasi biasa digunakan sebagai alat untuk melakukan prediksi dari data set yang terkumpul dalam satu atau beberapa kategori yang memiliki berbagai macam atribut. Sebelum masuk ke dalam tahap klasifikasi harus melewati beberapa tahap terlebih dahulu seperti pengumpulan data set jika belum ada, labeling data, *preprocessing* (pembersihan data), pembagian data menjadi data latih dan data uji serta membentuk model pembelajaran dari satu atau beberapa algoritma dalam data mining (Faizal, 2016).

2.7. Resampling Data

Resampling Data adalah sebuah teknik yang berguna untuk melakukan manipulasi data latih yang ada. Hal ini berguna untuk membuat model klasifikasi menjadi lebih baik. Data latih memiliki beberapa kekurangan seperti kecondongan dari distribusi kelas yang ada sehingga menimbulkan random undersampling dan juga random oversampling. Hal ini harus dilakukan perbaikan dengan metode resample agar memperbaiki masalah seperti random undersampling dan juga random oversampling yang terjadi sehingga membuat akurasi menjadi lebih baik

dari sebelum menggunakan teknik resample tersebut. Data yang diresample berasal dari dataset yang sudah ada dan membentuk data baru. Tujuan dari resampling data adalah untuk meningkatkan akurasi model klasifikasi (Zhang et al., 2011).

2.8. *Term Frequency Inverse Document Frequency*

TF-IDF merupakan kepanjangan dari *Term Frequency Inverse Document Frequency* yang memiliki arti sebagai teknik untuk melakukan pembobot kata dari dataset (Ariadi, & Fithriasari, 2015). Proses yang terjadi dalam *Term Frequency Inverse Document Frequency* adalah dengan cara melakukan perhitungan *Term Frequency* (TF) dengan juga IDF (*Inverse Document Frequency*). Proses TF-IDF berguna untuk menemukan kata yang sering muncul kemudian dihitung (TF), setelah kata dihitung berapa kali muncul maka selanjutnya dikalikan dengan berapa banyak dokumen dimana data tersebut muncul (IDF). Berikut ini rumus dalam melakukan pembobotan kata.

$$w_{ij} = tf_{ij} \times idf_j, idf_j = \log\left(\frac{N}{df_j}\right) \quad (2.1)$$

Keterangan dari persamaan diatas sebagai berikut.

- w_{ij} = Merupakan bobot dari kata i pada dokumen ke j.
- N = Jumlah dari seluruh dokumen yang ada
- TF = Jumlah dari keseluruhan kata i pada dokumen yang ada
- IDF = Jumlah dokumen j yang mengandung kata i

Proses TF-IDF dilakukan sebelum data masuk ke dalam tahap analisis menggunakan metode dalam data mining. Dalam penelitian ini menggunakan metode Naïve Bayes dan *Decision tree*.

2.9. Naïve Bayes

Naive Bayes menjadi salah satu metode klasifikasi dalam data mining atau text mining. Metode ini masuk ke dalam klasifikasi dalam melakukan prediksi secara sederhana. Berbagai kumpulan dari probabilitas yang ada akan dihitung dengan cara menjumlahkan kombinasi nilai dan juga frekuensi yang ada dalam suatu data set yang sudah dikumpulkan. Dalam metode Naïve Bayes setiap atribut yang ada dianggap bersifat independen yaitu tidak memiliki hubungan keterkaitan satu sama dengan yang lain (Nafalski & Wibawa, 2016). Adapun kelebihan yang bisa didapatkan dengan menggunakan metode Naïve Bayes yaitu dalam proses klasifikasi tidak perlu menggunakan data set yang besar. Sebelum melakukan klasifikasi data set dibagi menjadi data latih dan data uji. Dalam metode ini, data latih bisa diambil dengan jumlah yang sedikit. Tahap selanjutnya yaitu menentukan parameter varians dan mean yang diperlukan untuk proses klasifikasi dengan metode Naïve Bayes (Devita et al., 2018).

Dalam membentuk suatu model klasifikasi maka dataset yang sudah dibagi menjadi data uji dan data latih akan diberi label sesuatu dengan tujuan dari penelitian. Data yang sudah diberi label atau kategori akan melalui tahap proses untuk menghasilkan sebuah pengetahuan tentang nilai probabilitas dari setiap kata yang ada di dalam dokumen. Kata yang dihasilkan akan mewakili dokumen

dalam kategori tersebut. Adapun perhitungan yang digunakan untuk mencari probabilitas kata muncul dari setiap kategori yang ada dapat dilihat pada persamaan berikut.

$$(w_i | c_j) = \frac{1+n_i}{n+|kosakata|} \quad (2.2)$$

dimana:

- $(w_i | c_j)$: probabilitas kata pada setiap kategori
 n_i : frekuensi kemunculan kata setiap kategori
 n : jumlah seluruh kata dalam dokumen pada kategori tertentu
 $|kosakata|$: jumlah total kata di semua data latih

$$(c_j) = \frac{n(doc_j)}{n(sampel)} \quad (2.3)$$

dimana:

- (c_j) : probabilitas dokumen kategori
 (doc_j) : jumlah seluruh dokumen pada suatu kategori
 $n(sampel)$: jumlah seluruh dokumen latih

Data yang sudah melewati proses pelatihan untuk pembentukan model akan melalui tahap proses klasifikasi dimana data uji yang belum diketahui kategori atau label akan diprediksi sesuai dengan kata unik yang sudah didapat dari setiap kategori atau label $(w_i | c_j)$. Selanjutnya, (c_j) atau probabilitas setiap dokumen akan dihitung dari pengetahuan yang sudah tersimpan pada tahap pembuatan model klasifikasi. Adapun persamaan yang digunakan seperti berikut.

$$(c_j) \prod (w_i | c_j) \quad (2.4)$$

Langkah selanjutnya dari metode Naïve Bayes adalah membandingkan dan mencari nilai probabilitas maksimum $cMAP$ yang digunakan untuk mengklasifikasikan data uji pada dokumen berbahasa Indonesia yang akan diklasifikasikan ke dalam salah satu kategori yang tersedia. (Schneider, 2005), perhitungan tersebut dapat dilihat pada Persamaan 2.5.

$$cMAP = \operatorname{argmax}_{c_j \in c} (c_j) \prod_i (w_i | c_j) \quad (2.5)$$

2.10. Decision Tree

Pohon keputusan adalah representasi sederhana untuk mengklasifikasikan contoh. Proses dalam pohon keputusan adalah mengubah data (tabel) menjadi model pohon dan kemudian mengubah model pohon menjadi aturan. Algoritma yang digunakan adalah algoritma ID3. Algoritma ID3 menggunakan konsep entropy dan information gain. Nilai entropi dapat ditemukan menggunakan rumus

$$\operatorname{Entropy}(S) = \sum_{j=1}^k -P_j \log_2 P_j \quad (2.6)$$

Untuk nilai *Information Gain* (IG) dapat ditemukan dengan menggunakan rumus berikut ini :

$$\operatorname{Gain}(A) = \operatorname{Entropy}(S) - \sum_{j=1}^k \frac{|S_j|}{|S|} \times \operatorname{Entropy}(S_j) \quad (2.7)$$

2.11. Pengukuran Performa Klasifikasi

Pada analisis sentimen, tahap akhir yang dilakukan adalah evaluasi dari model klasifikasi yang sudah dibuat. Tahap evaluasi ini yaitu tahap untuk mengukur performa dari model klasifikasi yang sudah dibuat atau dibangun. Dalam evaluasi ada beberapa yang dihitung yaitu akurasi model klasifikasi, *recall*, *precision* dan

F-score (Hotho, Nurnberger, & Paass, 2005). Akurasi memiliki arti sebagai persen dari total dokumen yang diprediksi secara tepat dari data uji yang sudah dibagi dalam tahap pembagian data. *Recall* memiliki arti sebagai bagian kecil dari dataset yang relevan untuk diambil. Adapun persamaan dalam tahap evaluasi yang digunakan sebagai berikut.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.8)$$

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (2.9)$$

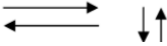



$$\text{Recall} = \frac{TP}{TP+FN} \quad (2.10)$$






$$F = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (2.11)$$

2.12. Diagram Alir (*Flowchart*)

Pada penelitian (Jogiyanto, 2015), Flow Chart dapat diartikan sebagai alur atau bagan yang dapat menunjukkan tahapan di dalam prosedur sistem secara logika atau di dalam program. Berikut ini Tabel 2.1 yang berisi penjelasan mengenai beberapa simbol atau ikon yang digunakan dalam diagram alir atau flowchart.

Tabel 2.1 Simbol-simbol Yang digunakan dalam Flowchart

Simbol	Keterangan
	Menunjukkan alur data atau alur proses untuk memperlihatkan jejak perjalanan data.
	State diagram dalam program
	Diagram untuk mewakili input (masukan) atau output (keluaran)
	Ikon titik akhir menunjukkan awal dan akhir suatu proses

Simbol	Keterangan
	Menunjukkan pekerjaan manual
	Digunakan untuk mewakili suatu proses
	Digunakan untuk merujuk pada tautan yang rusak antar halaman
	Digunakan untuk menunjukkan koneksi yang terputus antara halaman
	Untuk rinciannya ditunjukkan di tempat lain

(Sumber : (Jogiyanto, 2015))

2.13. Penelitian Terkait

Tabel 2.2 Penelitian Terkait

No.	Penulis	Judul	Metode	Hasil
1	(Jati, 2020)	Analisis Sentimen Produk Iphone Menggunakan <i>Naïve Bayes Classifier</i>	<i>Naïve Bayes Classifier</i>	1. Metode <i>Naïve Bayes</i> dapat melakukan sentimen analisis terhadap komentar berupa text. 2. Metode <i>Naïve Bayes</i> mendapatkan nilai akurasi 94% dari data uji
2	(Ernamia & Herliana, 2022)	Analisis Sentimen Kuliah Daring dengan Algoritma <i>Naïve Bayes, K-NN</i> dan <i>Decision Tree</i>	<i>Naïve Bayes, K-NN dan Decision Tree</i>	Hasil dalam peneitian ini berupa pembuktian bahwa algoritma <i>Naïve Bayes</i> yang digunakan memiliki akurasi yang lebih tinggi dibanding <i>KNN</i> dan <i>Decision Tree</i> . Akurasi dengan <i>Naïve Bayes</i> yaitu 81.75%, akurasi <i>KNN</i> yaitu 62.10 dan akurasi <i>Decision Tree</i> yaitu 51.89%.
3	(Salmon Pattiha, 2022)	Perbandingan Metode <i>K-NN, Naïve Bayes,</i>	<i>K-NN, Naïve Bayes,</i>	Hasil dari penelitian ini menunjukkan bahwa metode <i>Naïve Baye (NB)</i> memiliki akurasi yang lebih tinggi

No.	Penulis	Judul	Metode	Hasil
		<i>Decision Tree</i> untuk Analisis Sentimen Tweet Twitter Terkait Opini Terhadap PT PAL Indonesia	<i>Decision Tree</i>	dibanding metode yang lain dengan nilai akurasi yaitu 84,08% dan standar deviasi(+/-) 1,28%
4	(Ramadhan & Wahyudin, 2022)	Analisis Sentimen Mengenai Keberhasilan Indonesia di Ajang Thomas Cup 2020 (Studi Kasus Media Sosial Twitter) Menggunakan Metode <i>Naïve Bayes</i> dan <i>Decision Tree</i>	<i>Naïve Bayes dan Decision Tree</i>	Pada penelitian ini, dapat diketahui bahwa hasil yang diperoleh dengan metode <i>Naïve Bayes</i> mencapai akurasi 83.00% +/-10.5% (micro average:83.00%) sedangkan untuk metode <i>Decision Tree</i> nilai ujinya mencapai akurasi 77.00% +/- 4.58% (micro average: 77.00%).
5	(Tri Romadloni et al., 2019)	Perbandingan Metode <i>Naïve Bayes</i> , <i>KNN</i> dan <i>Decision Tree</i> Terhadap Analisis Sentimen Transport KRL Commuter Line	<i>Naïve Bayes, KNN dan Decision Tree</i>	Hasil analisis menggunakan KRL dengan metode <i>Naïve Bayes</i> dengan akurasi 80,00%, presisi 66,67%, sensitivitas 100%, spesifisitas 66,67%. Sedangkan metode KKN memiliki akurasi 80%, presisi 100%, sensitivitas 50%, dan spesifisitas 100%. Terakhir, metode <i>Decision Tree</i> mencapai akurasi 100%, akurasi 100%, sensitivitas 100%, dan spesifisitas 100%.
6	(Jtik et al., 2022)	Analisis Sentimen Pembelajaran Daring menggunakan Metode, <i>Naïve Bayes</i> ,	<i>Naive Bayes, KNN dan Decision Tree</i>	Penelitian ini menggunakan tiga metode yaitu <i>Naïve Bayes</i> , <i>KNN</i> dan <i>Decision Trees</i> . Akurasi dari 3 metode tersebut masing-masing adalah 55,49%, 61,47% 61,92%. Hasil yang benar adalah 46,45%, 53,90 dan 73,63%. Hasil recovery

No.	Penulis	Judul	Metode	Hasil
		<i>KNN</i> dan <i>Decision Tree</i>		dari 3 metode adalah 58,91%, 40,07 dan 11,42%. Dari hasil tersebut dapat disimpulkan bahwa metode pohon keputusan memiliki tingkat akurasi yang paling tinggi diantara kedua metode lainnya.
7	(Mujahidin et al., 2022)	Implementasi Analisis Sentimen Masyarakat Mengenai Kenaikan Harga BBM pada Komentar Youtube dengan Metode <i>Gaussian Naïve Bayes</i>	<i>Gaussian Naïve Bayes</i>	Berdasarkan hasil penelitian dapat disimpulkan bahwa menggunakan fungsi Fasttext sebagai vektorisasi kata dengan metode Bag of Word untuk membandingkan selisih akurasi yang diperoleh Gaussian Naïve Bayes. Fitur Fasttext mencapai hasil akurasi 0,74% sedangkan ekstraksi fitur Bag of Word mendapat skor rendah dengan kumpulan data yang sama 0,68%.
8	(Reynaldhi & Sibaroni, 2021)	Analisis Sentimen Review Film pada Twitter menggunakan Metode Klasifikasi Hybrid <i>SVM</i> , <i>Naïve Bayes</i> dan <i>Decision Tree</i>	<i>SVM</i> , <i>Naïve Bayes</i> dan <i>Decision Tree</i>	Pada penelitian ini, peneliti memperoleh hasil percobaan klasifikasi dengan menggunakan metode hybrid. Di mana pendekatan ini dapat memiliki kinerja yang lebih baik daripada pengklasifikasi mana pun yang bekerja secara individual. Adapun hasil yang didapat dari penelitian ini berupa skor akurasi sebesar 0.88, nilai recall sebesar 0.76, nilai F1-score yaitu 0.78 dan yang terakhir yaitu nilai presisi yaitu 0,8189.
9	(Hozairi et al., 2021)	Implementasi <i>Orange Data Mining</i> Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model <i>K-Nearest</i>	<i>KNN</i> , <i>Decision Tree</i> dan <i>Naïve Bayes</i>	Hasil penelitian ini memperoleh hasil kinerja <i>Naïve Bayes</i> , <i>KNN</i> dan <i>Decision Tree</i> . Dari 35 data uji yang digunakan <i>Naïve Bayes</i> memiliki nilai <i>Accuracy</i> sebesar 89%, <i>Precision</i> 76%, <i>KNN</i> memiliki nilai <i>Accracy</i> sebesar 77%, <i>Precision</i> 76%

No.	Penulis	Judul	Metode	Hasil
		<i>Neighbor, Decision Tree Serta Naïve Bayes</i>		dan <i>Decision Tree</i> memiliki nilai <i>Accuracy</i> 74% dan <i>Precision</i> 84%
10	(Rozaq et al., 2022)	Analisis Sentimen Terhadap Implementasi Program Merdeka Belajar Kampus Merdeka Menggunakan <i>Naïve Bayes, K-Nearest Neighbors</i> dan <i>Decision Tree</i>	<i>Naïve Bayes, KNN dan Decision Tree</i>	Pada penelitian ini dapat dijelaskan bahwa data hasil penemuan dibagi menjadi tiga kelas yaitu positif, negatif dan netral dengan metode <i>Naïve Bayes, KNN</i> dan <i>Decision Tree</i> sebanyak 475 data dibagi menjadi 2, untuk data latih dan data uji. Data pelatihan adalah 80% dari total data dan 20% sisanya digunakan sebagai data uji dengan akurasi 99,22% untuk <i>Naïve Bayes, KNN</i> 96,90% dan pohon keputusan 37,21%.

2.14. Matriks Penelitian

Tabel 2.3 Matriks Penelitian

No	Penulis	Algoritma		
		<i>KNN</i>	<i>SVM</i>	<i>Gaussian Naïve Bayes</i>
1.	(Jati, 2020)			
2.	(Ernamia & Herliana, 2022)	✓		
3.	(Salmon Pattiha, 2022)	✓		
4.	(Ramadhan & Wahyudin, 2022)			
5.	(Tri Romadloni et al., 2019)	✓		
6.	(Jtik et al., 2022)	✓		
7.	(Mujahidin et al., 2022)			✓
8.	(Reynaldhi & Sibaroni, 2021)		✓	
9.	(Hozairi et al., 2021)	✓		
10.	(Rozaq et al., 2022)	✓		