BAB II

TINJAUAN PUSTAKA

2.1 Analisis Sentimen

Ekskavasi padangan atau dapat disebut menjadi analisis sentimen merupakan tahapan mekanis untuk mengkaji komentar yang mampu mengetahui intonasi emosi, seperti pendapat negatif, positif, hingga netral dalam tanggapan masyarakat. Mekanisme ini termasuk dalam *Natural Language Processing* (NLP) yang mencakup pengaplikasian *machine learning, data mining, artificial intelligence*, dan tatabahasa komputasi untuk *text mining* bertujuan memperoleh kecenderungan pendapat dan informasi bersifat subjektif. Analisis sentimen umumnya digunakan untuk menemukan dan membedakan kondisi afektif dan informasi subjektif dari teks elektronik. Hal ini diimplementasikan diberagam tipe data tekstual, seperti opini pengguna di *social media*, materi perawatan kesehatan, serta dalam macam-macam konteks, seperti perniagaan sampai layanan pelanggan tujuan klinis (Mayur Wankhade, 2022).

Secara keseluruhan, proses ini dapat digunakan untuk mengumpulkan data dari pendapat atau sentimen yang kemudian dikategorikan sebagai sentimen positif dan negatif dari suatu media. Dalam penelitian ini, media yang akan dipakai bersumber dari Google Play Store dan App Store.

2.2 *Text Mining*

Text mining adalah subbagian dari *data mining* dan juga dikenal sebagai data text mining atau penemuan pengetahuan dalam teks. Proses ini memperoleh berupa pola dan tren yang ada di data tidak terstruktur, terutama teks, dengan menerapkan pembelajaran mesin, statistik, dan linguistik. Analisis teks dapat menghasilkan wawasan yang lebih kuantitatif dengan mengorganisir data. Pengelompokan teks, ekstraksi konsep, analisis sentimen, dan peringkasan adalah beberapa manfaat dari text mining. Hal

ini dapat membantu menjawab pertanyaan penelitian tentang teks dalam jumlah besar yang tidak mungkin atau tidak mudah diakses dengan teknik tradisional. (Aaryan Gupta, 2020).

2.3 Segari

Segari merupakan perusahaan yang bergerak di bidang grosir bahan makanan yang diresmikan pada tahun 2020 yang sedang bertumbuh pesat di Indonesia seperti di daerah Jakarta, Tangerang, dan Bekasi. Dengan menyediakan berbagai macam kebutuhan pokok seperti berbagai macam sayuran, buah-buahan, dan kebutuhan lainnya. Segari diluncurkan oleh Yosua Setiawan, Farandi Ramadhana, dan Farand Anugerah, Segari ditujukan untuk dapat membantu para petani tradisional. CEO Segari Yosua Setiawan mengatakan bahwa "rantai distribusi pertanian merupakan salah satu masalah yang paling kompleks di Indonesia." Masih banyak lapisan dari petani hingga produk pertanian sampai ke tangan konsumen. Kami berharap dapat memberikan dampak positif bagi konsumen dimana mereka dapat menerima makanan yang berkualitas dengan lebih cepat dan dari sisi petani, kami juga dapat memberikan harga yang adil untuk produk yang mereka jual (Derivanti & Aulia, 2023).

2.4 *Text Preprocessing*

Preprocessing teks merupakan tahapan memelihara dan memodifikasi data mentah menjadi format yang mampu dianalisis oleh algoritma *machine learning* dan . *Natural Language Processing* (NLP). Ketentuan dalam melakukan *pre-processing* bervariasi tergantung pada jenis korpus dan jenis proses pengolahan bahasa natural yang ingin digunakan para peneliti untuk menganalisis data. Metode *pre-processing* teks yang sering digunakan ialah menghapus format, normalisasi teks, tokenisasi, menanggulangi tanda baca, stemming, menghapus stopwords, dan menemukan ekspresi yang memiliki lebih dari satu kata. *Pre-processing* teks dilakukan dengan maksud untuk menyediakan korpus bagi pemodelan

serta menaikkan akurasi tugas-tugas NLP seperti analisis sentimen, pengenalan entitas bernama, dan pemodelan topik. Berbagai bidang, seperti ilmu kesehatan, penelitian organisasi, dan pengajaran mesin, menggunakan preprocessing teks. (Hickman, Thapa, Tay, Cao, & Srinivasan, 2020).

2.4.1 *Case Folding*

C*ase folding* merupan tahapan memetakan *string* ke bentuk yang menghilangkan huruf besar dan kecil. Hal ini bertujuan untuk pembanding teks tanpa huruf besar-kecil, seperti pengidentifikasi dalam program komputer, daripada transformasi teks yang sebenarnya. Oleh karena itu, proses ini sangat berperan penting untuk menyederhanakan tahapan pencarian kata dengan mengubah kata ke bentuk awal. (Alita & Rahman, 2020).

2.4.2 *Data Cleaning*

Data cleaning ialah tahapan menemukan dan memeriksa anomali, ketidakakuratan dan ketidakkonsistenan dalam *dataset* untuk menaikkan kualitas dari data yang digunakan. Proses ini termasuk menemukan dan memperbaiki data yang berlebihan, memvalidasi nilai variabel secara logis, memperlakukan pencilan, dan memproses data yang hilang. *Cleaning data* terdiri dari siklus penyaringan, diagnosis, dan pengeditan data yang diragukan. Pembersihan data sangat penting untuk menjamin kualitas dan validitas penelitian. (Broeck, Cunningham, Eeckels, & Herbst, 2005).

2.4.3 *Stemming*

Stemming merupakan tahapan menurunkan kata berimbuhan atau turunan ke bentuk kata dasar, pangkal, atau akarnya. Dalam proses ini, akhiran kata atau imbuhan lain dihilangkan untuk memastikan bahwa kata dengan arti yang sama terhubung ke stem yang sama, meskipun stem tersebut bukan akar yang sah. *Stemming* sangat penting untuk pemahaman bahasa alami, pengolahan bahasa

alami, dan studi morfologi linguistik. Ini juga penting untuk pekerjaan seperti penggalian teks dan pencarian informasi. Metode ini bertujuan untuk menaikkan *performance* sistem dalam memproses penemuan informasi dengan menangani kata-kata dengan stem yang sama sebagai sinonim, meningkatkan cakupan pertanyaan pencarian. (Wahyudi, Susyanto, & Nugroho, 2017).

2.4.4 *Normalization*

Normalization adalah bagian penting dari pembersihan data dan jaminan kualitas karena membantu menjaga integritas data dan memastikan bahwa data dapat dianalisis dan diinterpretasikan dengan baik. Proses ini melibatkan identifikasi dan koreksi kesalahan, ketidakkonsistenan, dan ketidaktepatan ejaan kata untuk meningkatkan kualitas data. (Finansyah, Afiahayati, & Sutanto, 2022).

2.4.5 *Stopword*

Kata-kata umum seperti "yang", "dan", "atau", dan lain-lain dihilangkan dari teks melalui teknik preprocessing teks yang dikenal sebagai *stopword removal*. Untuk meningkatkan kinerja sistem dengan berkonsentrasi pada kata-kata yang lebih bermakna dalam teks, proses ini sangat penting dalam proses pemrosesan bahasa natural (NLP) dan tugas-tugas pemrosesan teks. Namun, perlu diingat bahwa efek penghapusan stopword dapat berbeda-beda tergantung pada tugas tertentu; dalam beberapa kasus, ini telah terbukti mengganggu kinerja (misalnya, mendeteksi plagiarisme). (Ladani & Desai, 2020).

2.5 *Indonesian sentiment Lexicon*

Kamus *Indonesian Sentiment Lexicon*, juga disebut *InSet Lexicon*, digunakan untuk menemukan komentar atau pendapat. Kamus ini mengategorikan komentar menjadi opini posisif, netral, atau negatif. Cara kerja dari metode ini yaitu dengan menjumlahkan nilai masing-masing kata pada suatu kalimat, jika jumlahnya >0 maka akan dikelompokkan sebagai kalimat positif, jika jumlah nilainya 0 maka akan dikelompokkan menjadi kalimat netral, namun jika jumlah nilainya <0 makan akan dikelompokkan sebagai kalimat negatif (Musfiroh, Khaira, Utomo, & Suratno, 2021).

2.6 Pembobotan Kata

Untuk memungkinkan pemodelan, setelah *preprocessing* dan labeling teks selesai, data teks harus diberi pembobotan atau berat. Proses pembobotan kata akan didukung oleh metode *Frequency-Inverse Term Document*, atau TF-IDF. TF-IDF terdiri dari dua bagian: TF dan IDF. (Lavin, 2019).

2.6.1. TF (*Term Frequency*)

Term Frequency yang terdiri dari kata, frasa, atau elemen *indexing* yang muncul pada dokumen disebut "frekuensi", dan frekuensi munculnya sebuah *term* pada dokumen sehubungan dengan bobotnya. Persamaan TF untuk persamaan di bawah ini adalah sebagai berikut:

$$tf(t,d) = \frac{\text{jumlah kata } t \text{ dalam dokumen } d}{\text{total kata dalam dokumen } d}$$

2.6.2. IDF (*Inverse Document Frequency*)

Inverse Document Frequency diimplementasikan untuk menurunkan berat satuan *term* di berbagai dokumen yang sering dilihat. Karena keadaan ini, istilah ini adalah yang paling sering muncul di berbagai dokumen. Karena itu, itu adalah istilah jamak,

sehingga nilainya tidak terlalu penting. Persamaan IDF untuk persamaan di bawah ini: (Manurung, Matondang, & Prasvita, 2022).

$$\text{IDF}(t, D) = \log\left(\frac{\text{jumlah total dokumen dalam korpus } D}{\text{jumlah dokumen yang mengandung kata } t + 1}\right)$$

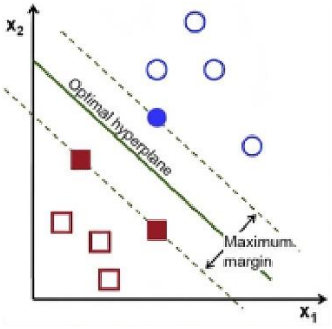
2.6.3. TF-IDF

Term Frequency – Inverse Document Frequency didapat dengan cara mengkalikan nilai TF dan IDF sebelumnya sudah didapat.

TF-IDF*(t,d,D)* = TF*(t,d) x* IDF*(t,D)*

2.7 *Support Vector Machine*

Model *supervised leaning* yang digunakan untuk klasifikasi dan analisis regresi adalah *Support Vector Machine* (SVM). *Hyperplane* adalah fungsi yang memiliki kemampuan untuk membedakan antara dua kelas. SVM sendiri bekerja untuk menemukan *hyperplane* maksimal. SVM akan memaksimalkan jarak antara pola pelatihan dan batas keputusan selama proses. Beberapa keunggulan algoritma ini termasuk kinerjanya yang luar biasa baik untuk jumlah data yang kecil maupun besar, dan kinerjanya yang luar biasa pada data yang memiliki banyak atribut dan mudah digunakan. Berikut adalah contoh SVM yang menggunakan *hyperlane* atau garis pembatas terbaik sebagai pemisahan antar dua kelas. Adapaun kernel yang dapat digunakan dalam pemodelan SVM, *linear* kernel dipakai untuk data yang *linier separable*, *polynomial* kernel untuk data yang memiliki *hubungan non-linier polynomial*, dan *radial base function* (RBF) untuk data yang memiliki hubungan *non-linier* kompleks (Abdusyukur, 2023).



Gambar 2. 1 Illustrasi Metode *Support Vector Machine*

(sumber: (Muhammad, 2022))

Untuk menemukan titik maksimal, garis *hyperplane* optimal dapat digunakan untuk memisahkan dua kelas. Untuk mendapatkan *hyperplane* SVM, persamaan dapat digunakan:

$(w \,.\, x_i) + b = 0$

Untuk data xi kelas -1 atau negatif, persamaan dapat dirumuskan:

$(w \,.\, x_i + b) \leq 1 \,,\, y_i = -1$

Namun, data xi dari kelas +1 atau kelas positif dapat dirumuskan dengan persamaan:

$(w \,.\, x_i + b) \geq 1 \,,\, y_i = 1$

w adalah vektor bobot (koefisien).

x adalah vektor fitur (representasi numerik dari data).

b adalah bias atau *intercept*.

Untuk menilai hasil klasifikasi dengan data uji, perlu menggunakan matriks *confusion* untuk mendukung proses ini. Hasil dari proses ini mencakup nilai akurasi, presisi, dan recall untuk klasifikasi yang telah dilakukan sebelumnya.

Tabel 2. 1 *Confusion Matrix*

		Nilai Aktual	
		Positif	Negatif
Nilai	Positif	TP	FP

	Negatif	FN	TN

Berikut persamaan untuk mencari hasil evaluasi (Handul, Matulessy, & Kaesmetan, 2024).

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Presisi kelas positif} = \frac{TP}{TP+FP}$$

$$\text{Presisi kelas negatif} = \frac{TN}{TN+FP}$$

$$\textit{Recall} \text{ kelas positif} = \frac{TP}{TP+FN}$$

$$\textit{Recall} \text{ kelas positif} = \frac{TN}{TN+FN}$$

$$\textit{F1-Score} = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision+Recall}$$

2.8 Python

Python adalah bahasa pemrograman yang populer yang digunakan untuk analisis sentimen, yang merupakan teknik untuk mengidentifikasi nada emosional teks. Beberapa librari Python yang mendukung analisis sentimen termasuk *scikit-learn*, *Inset Lexicon*, dan NLTK (*Natural Language Toolkit*). Untuk analisis sentimen, fitur seperti tokenisasi, penandaan bagian ucapan, dan pengenalan entitas bernama diberikan oleh library ini. Sangat cocok untuk tugas analisis sentimen karena mudah digunakan dan didukung oleh banyak library. *Sentiment Analysis and Cognition Engine* ialah alat yang ditulis dalam bahasa Python yang digunakan untuk pemrosesan teks dengan menggunakan vektor kata yang telah ditentukan sebelumnya dari berbagai basis data sumber untuk tujuan analisis sentimen terutama dalam konteks narasi klinis. (Mahawardana, Sasmita, & Pratama, 2022).

2.9 Penelitian Terdahulu

Berikut adalah beberapa referensi dari penelitian terdahulu yang menjadi bantuan bagi penulis dalam melakukan penelitian ini.

Tabel 2. 2 Penelitian Terdahulu

No.	Nama Peneliti	Judul Penelitian	Metode	Hasil Penelitian
1.	Fachran Sandi, (2023)	Klasifikasi Ulasan Pengguna Menggunakan Metode *Support Vector Machine* Pada Aplikasi Halodoc	*Support Vector Machine*	Penulis menggunakan teknik *data scraping* yang diambil dari ulasan pengguna aplikasi Halodoc versi 10 yang diambil pada periode Juli sampai dengan November 2021 sebanyak 880 data ulasan, sebanyak 448 ulasan menunjukan pengguna tidak suka dan 432 ulasan pengguna suka dengan versi Halodoc 10 ini. Algoritma SVM dapat mengklasifikasi dengan nilai akurasi sebesar 96,02% dengan kernel *linear* sedangkan dengan kernel *sigmoid* didapatkan nilai akurasi sebesar 78,97%.
2.	Rohanda Selia Pangaribuan, (2023)	Analisis Sentimen Pada Ulasan Pengguna Aplikasi Shopee di Google Play Store Menggunakan Algoritma *Support Vector Machine*	*Support Vector Machine*	Penulis menggunakan metode *scraping* dengan g*oogle play scraper* untuk mengambil ulasan pengguna Shopee sebanyak 1327 data yang ada pada bulan Januari 2023 dengan sentimen negatif sebanyak 504 data dan 823 data sentimen positif. Algoritma SVM memperoleh akurasi sebesar 79,16% dengan bantuan kernel RBF.
3.	Abitdavy Athallah	Analisis Sentimen Pengguna Aplikasi Dana Berdasarkan	*Support Vector Machine*	Penulis menggunakan teknik *scraping* menggunakan *google play scraper* pada *Google Play*

	Muhammad, (2022)	Ulasan Pada *Google Play Store* Menggunakan Metode *Support Vector Machine*		*Store* yang dilakukan pada 21 November dengan mengambil 1366 ulasan aplikasi DANA dan didapatkan kelas data negatif sebanyak 483 dan data positif sebanyak 883. Dengan algoritma SVM penulis mendapatkan nilai akurasi sebesar 89,41% dengan menggunakan seleksi fitur *Chi Square.*
4.	Alfio Kusuma, (2022)	Analisis Sentimen Pada Ulasan Aplikasi Indodax di *Google Play Store* Menggunakan Metode *Support Vector Machine*	*Support Vector Machine*	Penulis menggunakan teknik *web scraping* pada *review* aplikasi Indodax dengan *google play scraper*. Data yang dikumpulkan sebanyak 1138 *review* pada periode Oktober 2021 dan didapatkan data sentimen negatif sebesar 573 dan sentimen positif sebesar 565. Penulis menggunakan metode SVM dengan tiga rasio perbandingan, yakni 60:40, 70:30, dan 80:20 dari tiga rasio ini didapat akurasi tertinggi dari rasio perbandaingan 80:20, yakni 85%.
5.	Daniel Dwi Eryanto Manurung, (2022)	Analisis Sentimen Pada Ulasan Aplikasi Jakarta Terkini (JAKI) di *Google Play Store* Menggunakan Metode *Support Vector Machine*	*Support Vector Machine*	Penulis menggunakan metode *google play scraper* untuk menunjang pengumpulan data review aplikasi Jakarta Terkini (JAKI) di *Google Play Store.* Didapatkan 1000 data yang dikumpulkan dari *review* periode 26 November 2019–15 Maret 2022. Dari data tersebut didapat 453 data positif dan 547 negatif. Dengan metode SVM untuk membantu klasifikasi, penulis mendapatkan akurasi

				sebesar 97%, dengan nilai *precision* 100%, *recall* 93,6%, dan *specificity* 100%.