BAB II
LANDASAN TEORI

2.1 Gramedia Digital

Gambar 2.1 Logo Gramedia Digital

Sejak diakuisisi oleh PT Gramedia Digital Nusantara pada akhir 2016, Scoop mencatatkan perkembangan bisnis yang signifikan. Salah satu indikasinya, peningkatan jumlah pembaca digital hingga 50% selama setahun terakhir. Kelvin Wijaya, Managing Director Gramedia Digital Nusantara mengatakan transformasi baru ini merupakan upaya untuk memaksimalkan bisnis dan pelayanan aplikasi bagi 4 juta penggunanya. "Selain itu, perubahan nama ini menjadi langkah untuk mengintegrasikan Gramedia Digital dengan situs Gramedia.com yang menjual buku-buku dengan format cetak," jelas Kelvin (Rafie, 2018).

2.2 Sentimen Analisis

Menurut Rahmat Burhanudin, sentimen analisis adalah perkembangan kontekstual teks yang mengidentifikasi dan mengekstrak informasi subjektif dalam sumber dan membantu para pebisnis untuk memahami sentimen sosial dari merek, produk atau layanan mereka saat memantau percakapan online (Burhanudin, 2018). Menurut Tsalis Annisa, proses penggunaan *text analytics* untuk mendapatkan

berbagai sumber data dari internet dan beragam platform media sosial (Annisa, 2019). Menurut Muhammad Dani, proses dari berbagai data berupa pandangan atau opini sehingga dihasilkan kesimpulan dari berbagai opini yang ada (Dani, 2018). Menurut Ghulam Asrofi Buntoro, analisis sentiment merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini (Buntoro, 2017).

2.3 Text Prerpocessing

Text Preprocessing merupakan proses penyusunan suatu teks menjadi data untuk dianalisis pada langkah selanjutnya (Maylawati & Saptawati, 2017). Pada umumnya, praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh system. Praproses sangat penting dalam melakukan analisis sentimen, terutama untuk media sosial yang Sebagian besar berisi kata – kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki *noise* yang besar (Mujilahwati, 2016). *Text Preprocessing* pada penelitian ini terdiri dari beberapa tahapan, yaitu sebagai berikut.

1. Cleansing

Cleansing adalah suatu tahap di mana karakter maupun tanda baca yang tidak diperlukan dibuang dari teks. Contoh karakter yang dibuang adalah tanda seru, tanda tanya, koma dan titik (Hadna, Santosa, & Winarno, 2016). Contoh kalimat yang dimasukkan “Saya sangat menyukai aplikasi ini, tapi agsk lemot.” akan diiubah menjadi “Saya sangat menyukai aplikasi ini tapi agak lemot”.

2. Stopword Removal

Setelah *cleansing*, maka dilakukan tahap *stopword removal*, yaitu dengan menghapus kata – kata yang sangat umum (Jumeilah, 2017). Kata yang termasuk dalam *stopword* contohnya adalah yang, dan, di, itu, dengan, untuk, dari, dalam, akan, pada, ini, juga, saya, serta, adalah, bahwa, lain, kamu, dan masih banyak lagi (Jumeilah, 2017). Contoh kalimat yang telah diproses sebelumnya "Saya sangat menyukai aplikasi ini tapi agak lemot" akan dibubah menjadi "sangat menyukai agak lemot".

3. Tokenizing

Tahap *tokenizing* adalah tahap pemotongan *string* masukan berdasarkan kata – kata yang menyusunnya atau dengan kata lain pemecahan kalimat menjadi kata (Jumeilah, 2017). Yang umum dilakukan pada tahap *tokenizing* adalah memotong kata pada *white space* atau spasi. Pada tahap ini membagi urutan karakter menjadi kalimat dan kalimat menjadi *token* (Jumeilah, 2017). Contoh kalimat yang telah diproses pada tahap sebelumnya akan diubah menjadi "[sangat], [menyukai], [agak], [lemot]".

4. Model Bag-of-Words

Bag-of-words adalah sebuah model yang mempelajari sebuah kosakata dari seluruh dokumen, lalu memodelkan tiap dokumen dengan menghitung jumlah kemunculan setiap kata (Putri & Hendrowati, 2018). Jumlah kata yang sama akan

dijumlahkan menjadi satu array. Contoh kalimat yang telah diproses sebelumnya akan menjadi “[sangat][1], [menyukai][1], [agak][1], [lemot][1]”.

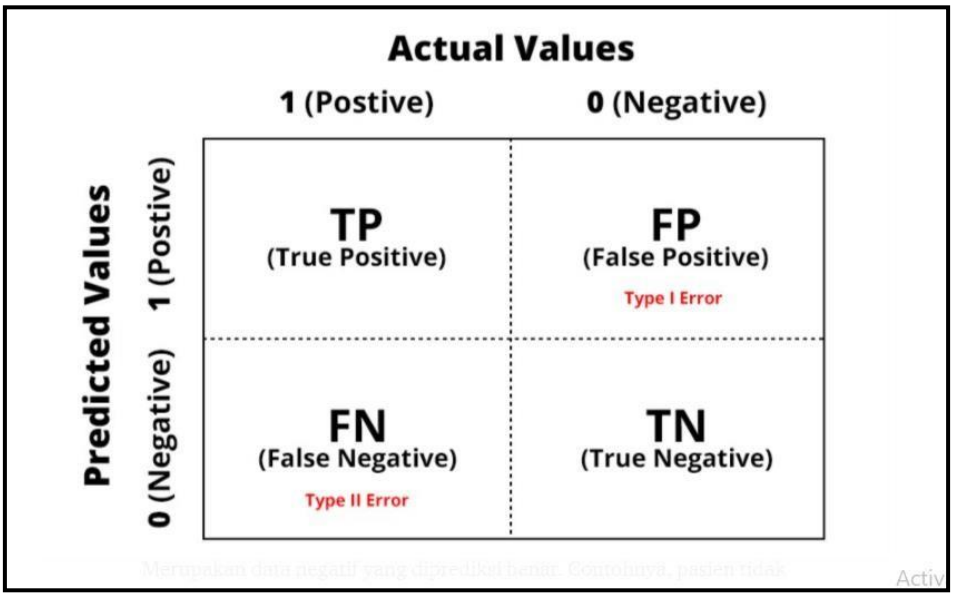
5. Term Frequency – Inverse Document Frequency

Term Frequency-Inverse Document adalah sebuah metode pembobotan yang menggabungkan dua konsep yaitu *Term Frequency* dan *Document Frequency*. *Term Frequency* adalah konsep pembobotan dengan mencari seberapa sering munculnya sebuah *term* dalam satu dokumen. Sedangkan *Dokumen Frequency* adalah banyaknya jumlah dokumen di mana sebuat term itu muncul (Hadna, Santosa, & Winarno, 2016). Merubah data yang telah diproses sebelumnya ke dalam bentuk vector agar dapat dibaca oleh sistem nantinya.

2.4 Confusion Matrix

Confusion matrix adalah suatu metode yang biasanya digunakan untuk melakukan perhitungan akurasi pada konsep *data mining*. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yamg benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan (Rahman, Darmawidjadja, & Alamsah, 2017). Pada dasarnya *confusion matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh sistem (model) dengan hasil klasifikasi sebenarnya (Nugroho, 2019).

Confusion matrix berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. *Confusion matrix* dapat digunakan dalam menghitung *performance metrics* untuk mengukur kinerja model yang telah dibuat. Terdapat beberapa *performance metrics popular* yang umum dan sering digunakan seperti *accuracy, precision, dan recall* (Nugroho, 2019).



Gambar 2.2 Confusion Matrix

Terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada *confusion matrix*. Keempat istilah tersebut adalah *True Positive* (TP) merupakan data positif yang diprediksi dengan benar, *True Negative* (TN) merupakan data negatif yang diprediksi benar, *False Positive* (FP) merupakan data negatif namun diprediksi sebagai data positif dan *False Negative* (FN) merupakan data positif namun diprediksi sebagai data negatif (Nugroho, 2019).

Ada cara yang lebih mudah untuk mengingatnya, yaitu:

- Jika diawali dengan ***True*** maka prediksinya adalah benar, diprediksi terjadi atau tidak terjadi.
- Jika diawali dengan ***False*** maka prediksinya adalah salah.
- Positif dan negatif merupakan hasil prediksi dari model

1. Accuracy

Accuracy menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. *Accuracy* merupakan tingkat kedekatan nilai prediksi dengan nilai sebenarnya (Nugroho, 2019).

$$accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100\% \quad \ldots (1)$$

2. Precision

Precision adalah keakuratan hasil klasifikasi dari seluruh dokumen oleh sistem, sehingga dapat diketahui apakah kategori data yang diklasifikasi sesuai dengan kategori yang sebenarnya. *Precision* dihitung dari jumlah pengenalan data yang bernilai benar oleh sistem dibagi dengan jumlah keseluruhan pengenalan data yang dilakukan pada sistem yang ditunjukkan rumus (Fauziah, Maududie, & Nuritha, 2018).

$$precision = \frac{TP}{TP+FP} \quad \ldots (2)$$

3. Recall

Recall menunjukkan tingkat keberhasilan sistem dalam mengenali suatu kategori. *Recall* dihitung dari jumlah pengenalan data yang bernilai benar oleh sistem dibagi dengan jumlah data yang seharusnya dapat dikenali sistem yang ditunjukkan dengan rumus (Fauziah, Maududie, & Nuritha, 2018).

$$recall = \frac{TP}{TP+FN} \quad \ldots (3)$$

4. F-measure

F-measure merupakan gambaran pengaruh relatif antara *precision* dan *recall* atau disebut *harmonic mean*. Performa algoritma yang digunakan dapat disimpulkan dari nilai *F-measure*. *F-measure* dapat dihitung seperti yang ditunjukkan dengan rumus (Fauziah, Maududie, & Nuritha, 2018).

$$F1 = 2 \times \frac{precision \times recall}{precision+recall} \quad \ldots (4)$$

2.5 Naive Bayes

Metode *Naive Bayes* merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema *Bayes* dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output (Manalu, Sianturi, & Manalu, 2017).

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad \ldots(5)$$

X : Data dengan *class* yang belum diketahui

H : Hipotesis data merupakan suatu *class* spesifik

P(H|X) : Probabilitas hipotesis *H* berdasar kondisi *X* (posteriori probabilitas)

P(H) : Probabilitas hipotesis *H* (prior probabilitas)

P(X|H) : Probabilitas *X* berdasarkan kondisi pada hipotesis *H*

P(X) : Probabilitas *X*

2.5 Multinomial Naive Bayes

Multinomial Naive Bayes merupakan sebuah metode yang bekerja dengan cara menghitung frekuensi setiap *term* pada dokumen (McCallum dan Nigam, 1998). Sebagai contoh, frekuensi kata "jaringan" pada berita teknologi. Sehingga peran tokenisasi dalam *Multinomial Naive Bayes* ini sangat penting. Dalam *Multinomial Naive Bayes*, dokumen urutan kejadian munculnya kata dalam dokumen tidak dipedulikan, jadi dokumen dianggap seperti "bag of words", sehingga setiap kata diolah menggunakan distribusi *multinomial* (Fanissa, Fauzi, & Adinugroho, 2018). Alasan menggunakan algoritma Multinomial Naive Bayes karena biasanya digunakan pada data rating karena setiap peringkat akan memiliki frekuensi tertentu dan memiliki hitungan setiap kata untuk memprediksi kelas.

Bernoulli Naive Bayes mengasumsikan semua fitur ke dalam bentuk biner sehingga hanya mengambil dua nilai yang mewakili kata tidak dan ya.