BAB 2

TINJAUAN PUSTAKA

2.1 Media Sosial

Media sosial adalah media online seperti blog, jejaring sosial, wiki, forum, dan dunia maya yang memungkinkan penggunanya dengan mudah berpartisipasi, berbagi, dan membuat konten yang terhubung melalui jaringan Internet (Istiani & Islamy, 2020). Perkembangan media sosial memungkinkan pengguna internet untuk terhubung dengan berbagai wilayah dan negara yang ditemuinya di dunia maya. Contoh aktivitas dunia maya antara lain *chatting*, mengirim pesan elektronik, dan berkomentar di *twitter* itulah beberapa momen yang ada di media sosial.

2.2 Twitter

Twitter merupakan media komunikasi paling sering digunakan hal ini dikarenakan *twitter* merupakan media sosial yang sederhana, mudah digunakan, dan memungkinkan pengguna bebas mengutarakan pandangan dan pendapatnya (Dedi Darwis, Eka Shintya Pratiwi, 2020). Hal tersebut menjadikan *twitter* menjadi salah satu media sosial untuk mendapatkan data teks. Data teks yang diperoleh diproses dengan cara yang sesuai untuk proses penambangan teks atau *text mining*.

2.3 Text Mining

Text mining atau disebut juga penambangan teks yaitu proses menemukan informasi dalam kumpulan teks dan *secara* otomatis mengidentifikasi pola dan hubungan yang menarik dalam data teks (Cartwright, 2010). *Text mining* mengekstrasi tinjauan opini dan analisis teks untuk membantu Anda memahami opini remaja dalam data teks (Kurniawan & Susanto, 2019). Tahapan utama dalam melakukan *text mining* ini dengan menemukan sekumpulan kata yang menggambarkan isi suatu dokumen, hubungan antar data atau dokumen dianalisis menggunakan metode statistika seperti klasifikasi, analisis kelompok, dan asosiasi.

2.4 Analisis Sentimen

Analisis sentimen adalah proses memahami dan mengolah data tekstual dengan tujuan memperoleh informasi tentang topik tertentu. Analisis sentimen merupakan proses menganalisis teks digital mengidentifikasi pandangan dan pendapat dalam teks terkait isu atau pokok bahasan, apakah cenderung positif atau negatif. Analisis sentimen sebenarnya merupakan proses klasifikasi yang tidak sesederhana proses klasifikasi biasa yaitu karena penggunaan bahasa dengan kata-kata yang ambigu, kurangnya intonasi dalam teks, dan perkembangan bahasa itu sendiri (Rahman, 2020).

2.5 RapidMiner

RapidMiner merupakan *open-source* terdepan di dunia kerangka kerja untuk penambangan informasi. RapidMiner menyediakan teknologi

penambangan data dan pembelajaran mesin seperti: Memuat dan mengonversi data. Pra proses dan visualisasi data, analisis prediktif dan pemodelan statistik, evaluasi dan penerapan. Rapidminer dibangun dengan bahasa pemrograman Java.RapidMiner menyediakan GUI untuk merancang dan menjalankan alur kerja analisis. Alur kerja ini disebut "proses" di RapidMiner dan terdiri dari sejumlah "operator". Setiap operator melakukan tugas dalam proses, dan output dari setiap operator membentuk input untuk operator berikutnya (Gupta, 2015).

2.6 Praproses Data

Praproses data dilakukan untuk membersihkan data yang belum siap digunakan agar dapat diproses pada langkah berikutnya. Tahap ini data yang tidak sesuai akan dihapus agar dapat diklasifikasikan. Pentingnya analisis sentimen terletak pada kenyataan bahwa media sosial penuh dengan kata-kata dan kalimat yang tidak dapat diatur atau biasa disebut dengan *noise*, sehingga menjadikan proses ini penting. Pada praproses data ada beberapa langkah, yaitu:

2.6.1 Tokenize

Tokenisasi adalah tahap pemisahan kata, simbol, frasa, dan entitas penting lainnya (disebut token) dari teks (Rahman, 2020).

2.6.2 Transform Cases

Transform cases adalah proses mengubah seluruh karakter pada data sesuai keinginan, seperti mengubah huruf besar menjadi huruf kecil dan sebaliknya (Maulana et al., 2020).

2.6.3 Filter Stopword

Pada tahap ini, semua kata diperiksa, jika dalam dokumen terdapat kata-kata yang tidak memberikan kontribusi banyak seperti kata sambung, kata depan, atau kata ganti, maka kata-kata tersebut akan dihilangkan (Maulana et al., 2020).

2.7 N-Gram

Setelah melalui tahap praproses data*(tokenize, transform data,* dan *filter stopword)* akan dilakukan proses n-gram pada dokumen teks. *N-gram* merupakan rangkaian *n* huruf yang berdekatan(termasuk tanda baca dan spasi), suku kata , atau kata utuh. *N-gram* menjelaskan bahwa bahasa tidak terdiri dari kata-kata tunggal, melainkan rangkaian kata-kata tunggal dan frase dua, tiga atau lebih kata, masing-masing berisi informasi unik (Fahrur Rozi et al., 2020). Contoh penerapan *n-gram* pada *bigram* atau n=2 yaitu "orang yang berbuat salah selalu merasa paranoid tapi cinta terkadang membuat kita tak tau mana yang salah dan benar" menjadi "orang yang", "yang berbuat", "berbuat salah", "salah selalu", "selalu merasa", "merasa paranoid", "paranoid tapi", "tapi cinta", "cinta terkadang", "terkadang membuat", "membuat kita", "kita tak", "tak tau", "tau mana", "mana yang",

"yang salah", "salah dan", "dan benar". Jika menggunakan *n-gram* pada *trigram* atau n-3 maka kalimat dipecah setiap 3 kata, jika terdapat tanda baca seperti koma atau titik koma, maka proses *n-gram* tidak akan melewatkannya dan akan membuat *n-gram* baru yang berisi kata setelah karakter tersebut.

2.8 Pembobotan Kata

Pembobotan kata atau menghitung skor didalam suatu dokumen disebut *Term Frequency-Inverse Document Frequency (TF-IDF).* Pembobotan TF-IDF adalah pembobotan kata atau term berdasarkan frekuensi kemunculannya (TF) dalam suatu dokumen. Persamaan frekuensi kata tf (t,d) ditunjukkan pada Persamaan berikut:

$$\text{tf (t,d)} = \frac{f_{t,d}}{\sum t\prime \in d f_{t',d}} \quad \text{.....(1)}$$

Keterangan:

tf (t,d) = jumlah kemunculan termtpada dokumen d.

$f_{t,d}$ = pencacahan mentah

$\frac{f_{t,d}}{\sum t\prime \in d f_{t',d}}$ = frekuensi istilah

Untuk menghitung IDF dapat menggunakan persamaan berikut:

$$\text{idf} = \log\left(\frac{N}{df}\right) \quad \text{.....(2)}$$

Keterangan:

idf = ukuran informasi yang diberikan oleh istilah *t*

N = Jumlah total dokumen yang digunakan.

df = Jumlah dokumen di mana term yang dipilih muncul.

Untuk menghitung TF-IDF dapat menggunakan persaman berikut:

$$\text{TF-IDF} = \text{tf (t,d)} \; x \; \text{idf} \quad \text{.....(3)}$$

Keterangan:

tf (t,d) : jumlah kemunculan termtpada dokumen d.

idf : ukuran informasi yang diberikan oleh istilah *t*

2.9 Klasifikasi Naïve Bayes

Menurut (Prasetyo, E. 2012) Klasifikasi adalah teknik menguji objek data dan mengidentifikasinya dalam satu kelas tertentu dari kumpulan kelas lain. Klasifikasi *Naïve Bayes* adalah Memprediksi peluang masa depan berdasarkan pengalaman masa lalu dikenal dengan teorema Bayes (Nugroho, 2018). Keuntungan menggunakan algoritma *Naïve Bayes* yaitu algoritma ini hanya membutuhkan data pelatihan kecil (Imron, n.d.).

Dibawah ini adalah rumus dari klasifikasi *Naïve Bayes*:

$$P(A|B) = \frac{P(A\backslash B)\ X\ P(A)}{P(B)} \quad \text{.....(4)}$$

Keterangan:

A: Data hipotetis berasal dari kelas tertentu.

B: Data dengan kelas tidak diketahui.

P(A|B): Probabilitas hipotesis berdasarkan kondisi.

P(A): probabilitas hipotesis.

P(B|A): Probabilitas berdasarkan kondisi hipotesis.

P(B): Probabilitas B.

Contoh perhitungan menggunakan *Naïve Bayes* menentukan *future customer:*

Tabel 2.1 Contoh dataset naïve bayes

age	gender	payment method	future customer?
old	male	credit card	yes
young	male	cheque	yes
young	female	credit card	yes
young	female	credit card	no
young	male	credit card	yes
young	female	cheque	no
young	female	credit card	yes
mid-age	male	credit card	yes
old	female	credit card	no
young	male	credit card	yes
young	male	credit card	yes
mid-age	female	cash	no
young	male	cash	yes
young	male	credit card	yes

P(yes) =10/14

P(no) = 4/14

P(old | yes) = 1/10

P(old | no) = 1/4

P(young | yes) = 7/10

P(young |no) =2/4

P(male | yes) = 8/10

P(male | no) = 0/4

P(female | yes) = 2/10

P(female | no) = 4/4

P(cheque | yes) =1/10

P(cheque | no) = 1/4

P(cash | yes) = 1/10

P(cash | no) = 1/4

P(credit | yes) = 8/10

P(credit | no) =2/4

- New Customer

  Age : old

  Gender : female

  Payment : cash

P(new | yes) = P(old | yes) * P(female | yes) * P(cash | yes) * P(yes)

P(new | yes) = 1/10 * 2/10 * 1/10 *10/14

P(new | yes) = 0.001423

P(new | no) = P(old | no) * P(female | no) * P(cash | no) * P(no)

P(new | no) = 1/4 * 4/4 * 1/4 * 4/14

P(new | no) = 0.017857

Dari contoh perhitungan *Naïve Bayes* diatas maka *new customer* bukan termasuk dalam *future customer* karena nilai dari P(new | no) > P (new | yes).

2.10 K-Fold Validation

Validasi yang digunakan dalam penelitian ini adalah *k-fold validation. K-fold validation* adalah metode memilah data menjadi *train* data dan *test* data (Ridwansyah, 2022). *K-fold validation* yang digunakan yaitu k=5 sampai dengan k=10, misal k yang digunakan adalah k=7 maka dataset dipecah menjadi 7 bagian, 6 bagian digunakan untuk data latih, dan 1 bagian digunakan untuk data uji. Proses ini dilakukan sebanyak 7 kali (Wahono et al., 2014). Dibawah ini merupakan tabel pembagian dataset dengan *k-fold validation.*

Tabel 2.2 Pembagian dataset *K-Fold Validation*

Validasi ke-n	Pembagian dataset						
1	■						
2		■					
3			■				
4				■			
5					■		
6						■	
7							■

Tanda hitam pada tabel 2.2 menandakan dataset yang digunakan untuk data uji. Bagian lainya menandakan dataset yang digunakan untuk data latih.

2.11 Pengukuran Performa

Pada tahap ini dilakukan pengukuran performa akurasi, presisi, dan *recall*. Pengukuran ini sering disebut dengan *confusion matrix*, *confusion matrix* adalah tabel yang menampilkan hitungan jawaban benar dan salah pada data tes. (Normawati & Prayogi, 2021), dengan *true positive (TP), true negative (TN), false positive (FP) dan false negative (FN)* sebagai indikator. Maka *confusion matrix*-nya dapat disajikan seperti pada tabel 2.3 dibawah ini

Tabel 2.3 *Confusion matrix*

Aktual Kelas	Terklasifikasi Positif (1)	Terklasifikasi Negatif (0)
Positif (1)	TP	FP
Negatif (0)	FN	TN

Keterangan:

- TP adalah *True Positive*, yaitu jumlah dokumen dari kelas 1 yang benar terklasifikasikan sebagai kelas 1
- TN adalah *True Negative*, yaitu jumlah dokumen dari kelas 0 yang benar terklasifikasikan sebagai kelas 0

- FN adalah *False Negative*, yaitu jumlah dokumen dari kelas 0 yang salah terklasifikasikan sebagai kelas 1
- FP adalah *False Positive*, yaitu jumlah dokumen dari kelas 1 yang salah terklasifikasikan sebagai kelas 0

  Rumus *confusion matrix* untuk menghitung akurasi, presisi, dan *recall* seperti berikut:

  Akurasi $= \frac{TP+TN}{TP+TN+FP+FN}$ .....(5)

  Presisi $= \frac{TP}{TP+FP}$ .....(6)

  Recall $= \frac{TP}{TP+FN}$ .....(7)

Contoh perhitungan menggunakan *Confusion Matrix*

Tabel 2.4 Contoh Confusion Matrix

Aktual Kelas	Terklasifikasi Positif (1)	Terklasifikasi Negatif (0)
Positif (1)	110	47
Negatif (0)	53	90

Pertama yang akan dicari adalah akurasi menggunakan persamaan nomor 5

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

Akurasi $= \frac{110+90}{110+90+47+53}$

Akurasi = 0.66666667

Kedua Persisi menggunakan persamaan nomor 6

Presisi $= \frac{TP}{TP+FP}$

Presisi $= \frac{110}{110+47}$

Persisi = 0.70063694

Terakhir menghitung *Recall* menggunakan persamaan nomor 7

Recall $= \frac{TP}{TP+FN}$

Recall $= \frac{110}{110+53}$

Recall = 0.67484663

Tabel 2.5 Hasil Contoh Confusion Matrix

Dataset	Akurasi	Presisi	Recal
300	0.66666667	0.70063694	0.67484663

Hasil dari contoh perhitungan di atas ditunjukkan di bawah ini. Berdasarkan hasil perhitungan *recall* dapat disimpulkan bahwa 67,48% data tergolong benar. Kumpulan data yang diprediksi bernilai negatif ternyata bernilai negatif, begitu pula sebaliknya. Nilai presisi tersebut kemudian digunakan untuk menentukan persentase prediksi negatif yang sebenarnya terhadap total hasil prediksi negatif, yaitu sebesar 70,06%. Tingkat akurasi model pada penelitian ini mencapai 66,66%. Artinya 66,66% kalimat diprediksi benar di semua kelas (kelas 0 atau sentimen negatif dan kelas 1 atau sentimen positif).