BAB II
LANDASAN TEORI

2.1. Analisis Sentimen

Salah satu bagian penting dalam pencarian informasi adalah mengetahui apa yang orang lain pikirkan, dan saat ini semakin banyak orang menyampaikan pikiran dan opini mereka secara bebas melalui internet kepada orang lain yang tidak mereka kenal (Pang & Lee, 2008). Teknologi informasi kini memberikan peluang untuk mengembangkan sistem yang dapat memahami opini orang lain secara otomatis, dan memberikan evaluasi mood pada opini seseorang di internet. Analisis mood pada opini disebut Analisis Sentimen, yang merujuk kepada analisis secara otomatis mengenai teks yang evaluatif dengan berfokus kepada klasifikasi teks berdasarkan polaritas yang dimiliknya (Pang & Lee, 2008). Klasifikasi data pada kelompok sentimen tertentu (positif atau negatif) dilakukan dengan membangun model probabilitas kemunculan suatu kata dalam dokumen yang telah dikelompokkan sebelumnya.

Besarnya pengaruh dan manfaat dari sentiment analysis, menyebabkan penelitian ataupun aplikasi mengenai sentiment analysis berkembang pesat, bahkan di Amerika ada kurang lebih 20-30 perusahaan menggunakan *sentiment analysis* untuk mendapatkan informasi tentang sentimen masyarakat terhadap pelayanan perusahaan (Sumartini, 2011). Pada dasarnya *sentiment analysis* merupakan klasifikasi, tetapi kenyataannya tidak semudah proses klasifikasi biasa karena terkait penggunaan bahasa. Terdapat ambigu dalam penggunaan kata, tidak adanya intonasi dalam sebuah teks, dan perkembangan dari bahasa itu sendiri (Pang & Lee, 2008).

2.2. Twitter

Twitter merupakan salah satu media sosial yang memungkinkan penggunanya untuk mengirim dan membaca pesan *tweet* berupa teks, gambar atau sebuah video. Media sosial twitter berbeda dengan media sosial lainnya terutama dalam penulisan status atau *tweet.* Media sosial selain twitter tidak ada batasan

karakter yang dapat dituliskan sedangkan twitter hanya memberikan 280 karakter yang dapat ditulis sebagai status atau cuitan.

Twitter bersifat publik, maksudnya adalah semua yang dituliskan atau dibagikan dapat dilihat oleh semua pengguna lainnya, namun penguna twitter dapat membatasi pengiriman *tweett* hanya bisa dilihat temannya saja atau biasa disebut sebagai *follower*. Twitter memiliki fitur utama yakni dapat menuliskan status atau cuitan serta dapat melakukan pengiriman pesan kepada pengguna lain, fitur lain dari sosial media twitter sebagai berikut :

1. ***Following***

   Salah satu fitur andalan dari media sosial twitter yakni *following*. Fitur ini memungkinkan pengguna untuk saling terhubung dengan pengguna lain atau bisa disebut sebagai pertemanan. Setiap unggahan *tweet* dari pengguna yang telah di *follow* maka dapat dilihat di beranda pengguna yang telah mengikutinya.

2. ***Retweet***

   Fitur *retweet* merupakan fitur yang memudahkan pengguna untuk meneruskan atau menyebarkan *tweet* pengguna lain sehingga dapat muncul di beranda pribadi

3. ***Hashtag***

   Hastag atau tagar merupakan fitur dari twitter yang dapat mengelompokan sebuah tweet. Dimana setiap *tweet* yang ditulis bisa ditambahkan dengan hastag berupa kata atau *keyword* dari *tweet* tersebut. Salah satu fungsi dari adanya *hastag* yakni untuk mengelompokan atau memudahkan pencarian terhadap kata kunci dari *tweet*.

4. ***Trending Topic***

   Fitur *tranding* topik merupakan fitur yang menampilkan topik atau hastag yang sedang popular atau banyak dibahas oleh pengguna *tweet*. Adanya *tranding topic* membuat pengguna mengetahui hal apa saja yang sedang *viral* di kalangan masyarakat.

2.3. Operator Seluler

Operator seluler atau operator nirkabel adalah perusahaan telepon yang menyediakan layanan untuk pengguna telepon seluler. Operator memberikan kartu SIM ke pelanggan yang memasukkan ke ponsel untuk mendapatkan akses ke layanan tersebut.

2.4. *Preprocessing*

Data *tweet* yang telah diambil dari sosial media twitter masih merupakan data mentah maka dari itu perlu dilakukan *preprocessing* untuk mendapatkan data yang siap untuk diproses selanjutnya. Dimana penjelasan dari tahap-tahap tersebut adalah sebagai berikut:

Tabel 2. *1* Contoh text *tweet*

Teks *input*
@gojekindonesia min, td saya go-send driver malah nganter paket ke jne padahal orderan ke wahana. Dan dia lepas tangan, jd gmn ya? :'(

2.4.1 *Cleasing*

Cleaning adalah membersihkan kalimat dari kata yang tidak diperlukan untuk mengurangi noise seperti karakter HTML, RT, ikon emosi, hashtag (#), username (@), url (http://situs.com), email (nama@situs.com), simbol dan tanda baca.

Tabel 2. *2* Contoh proses *cleasing*

Text *Output*
min td saya gosend driver malah nganter paket ke jne padahal orderan ke wahana Dan dia lepas tangan jd gmn ya

2.4.2 *Case Folding*

Dalam penulisan *tweet* sering ditemukan perbedaan bentuk huruf. Tahapan *case folding* akan merubah bentuk huruf menjadi huruf kecil atau disebut juga penyeragaman bentuk huruf.

Tabel 2. *3* Contoh proses *case folding*

Text *Output*
min td saya gosend driver malah nganter paket ke jne padahal orderan ke wahana dan dia lepas tangan jd gmn ya

2.4.3 *Tokenizing*

Tahap *tokenizing* yakni tahap pemotongan string inputan berdasarkan kata yang menyusunnya. Pada dasarnya proses *tokenizing* adalah pemenggalan kalimat menjadi kata.

Tabel 2. *4* Contoh proses *tokenizing*

Text *Output*
[min] [td] [saya] [gosend] [driver] [malah] [nganter] [paket] [ke] [jne] [padahal] [orderan] [ke] [wahana] [dan] [dia] [lepas] [tangan] [jd] [gmn] [ya]

2.4.4 *Stopword removal*

Stopword removal adalah proses menghilangkan kata-kata yang tidak memiliki arti seperti kata “yang”, “di”, “itu” dan lain sebagainya.

Tabel 2. *5* Contoh proses *stopword removal*

Text *Output*
min gosend driver anter paket jne orderan wahana lepas tangan

2.4.5 *Remove emoticon*

Remove emoticon adalah proses pengilangan *emoticon* pada *tweet*. Ketika sedang menulis status (*tweet*) seseorang kadang salah atau kurang tepat dalam penggunaan *emoticon*, entah disengaja atau tidak banyak yang melakukannya. Contohnya: Mereka hanya bisa memfitnah karena tidak bisa ketemu fakta buruk :), kata opini fitnah tapi emoticonnya senyum :), dengan begitu *emoticon* akan mengganggu dalam proses *Sentiment Analysis tweet*, jadi dalam proses ini *emoticon* dihapus atau diabaikan.

Tabel 2. *6* Contoh *Emoticon*

Emoticon	**Perasaan**	**Sentimen**
:) :-)	Senang	Positif
:(:-(	Sedih	Negatif
:D :-D	Sangat Senang	Positif
D: D=	Sangat Sedih	Negatif
_ *.* *-*	Kagum	Positif
D:< D: D8	Takut,jijik,kesedihan	Negatif
xD XD	Tertawa	Positif
:| =| :-|	Tanpa Expresi	Netral

Tabel 2. *7* Contoh proses *remove emoticon*

Text *Output*
min gosend driver anter paket jne orderan wahana lepas tangan

2.4.6 *Convert negation*

Convert negation dilakukan jika terdapat kata negasi sebelum kata yang bernilai positif, maka kata tersebut akan diubah nilainya menjadi negatif dan begitupun sebaliknya. Kata-kata yang bersifat negasi seperti “bukan”, “tidak”, “enggak”, “ga”, “jangan”, “nggak”, “tak”, dan “gak”.

Tabel 2. *8* Contoh proses *convert negation*

Sebelum Convert Negation	*Sesudah Convert Negation*
tidak bagus [negatif] x [positif]	tidakbagus [negatif]

2.4.7 *Stemming*

Stemming merupakan proses pengubahan kata menjadi kata dasar yang sesuai dengan aturan bahasa indonesia. Proses stemming menggunakan bantuan algoritma Sastrawi.

Tabel 2. *9* Contoh proses *stemming*

Text *Output*
min, gosend, driver, anter, paket, jne, order, wahana, lepas, tangan

2.4.8 Normalisasi

Normalisasi bertujuan mempermudah proses analisis sentimen terhadap entitasentitas dikarenakan banyaknya kata yang tidak baku didalam *tweet* seperti singkatan, tanggal, jumlah mata uang, dan akronim. Kata yang tidak baku memiliki kecenderungan yang lebih tinggi dalam hal ambiguitas interpretasinya atau pelafalannya dibanding kata yang sudah baku. Misalnya, kata betul bisa ditulis dengan kata btl, betol, bener, bnr, dan lain sebagainya. Jika tidak dilakukan normalisasi, kata-kata tersebut akan berdiri sendiri. Padahal, seharusnya kata-kata tersebut terkelompokkan ke dalam konteks yang sama, yaitu benar. Kasus semacam ini dapat diselesaikan dengan normalisasi teks dengan cara mengganti kata yang tidak baku dengan kata yang sesuai konteksnya (Sproat et al. 2001). Proses normalisasi teks dilakukan menggunakan penggantian pada kata yang tidak baku menjadi kata baku dengan sistem yang telah dibuat oleh (Aziz, 2013). Sebelum dilakukan penggantian dengan kata baku, harus dibuat terlebih dahulu sebuah kamus yang berisi kata yang tidak baku.

Tabel 2. *10* Contoh proses *normalisasi*

Text *Output*
admin, gosend, pengemudi, antar, paket, jne, order, wahana, lepas, tangan

2.5. *Term Frequency* (TF)

TF adalah frekuensi dari kemunculan sebuah term dalam dokumen yang bersangkutan. Semakin besar jumlah kemunculan suatu term (TF tinggi) dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar. Pada term *frequency* (TF), terdapat beberapa jenis formula yang dapat digunakan :

1. TF biner (binary TF), hanya memperhatikan apakah suatu kata atau term ada atau tidak dalam dokumen, jika ada diberi nilai satu (1), jika tidak diberi nilai nol (0).
2. TF murni (raw TF), nilai TF diberikan berdasarkan jumlah kemunculan suatu term di dokumen. Contohnya, jika muncul lima (5) kali maka kata tersebut akan bernilai lima (5).
3. TF logaritmik, hal ini untuk menghindari dominansi dokumen yang mengandung sedikit term dalam *query*, namun mempunyai frekuensi yang tinggi.
4. TF normalisasi, menggunakan perbandingan antara frekuensi sebuah term dengan nilai maksimum dari keseluruhan atau kumpulan frekuensi term yang ada pada suatu dokumen.

2.6. *Naïve Bayes Classifer* (NBC)

Salah satu tugas Data Mining adalah klasifikasi data, yaitu memetakan (mengklasifikasikan) data ke dalam satu atau beberapa kelas yaang sudah didefinisikan sebelumnya. Salah satu metoda dalam klasifikasi data adalah *Naïve Bayes Classifier* (NBC). Klasifikasi *bayes* yang ditemukan oleh Thomas *Bayes* paada abad ke-18. Teori klasifikasi *bayes* adalah pendekatan statistika yang fundamental dalam data mining. Pendekatan ini berdasarkan pada kuantifikasi trade-off antara berbagai keputusan klasifikasi dengan menggunakan probabilitas (Suyatno, 2017).

Dasar dari *Naïve Bayes* yang dipakai dalam pemrograman adalah rumus *Bayes*:

$$P(A|B) = (P(B|A) * P(A))/P(B) \quad (2.1)$$

Peluang kejadian A sebagai B ditentukan dari peluang B saat A, peluang A, dan peluang B. Pada pengaplikasiannya nanti rumus ini berubah menjadi :

$$P(\boldsymbol{C_i}|D) = (P(D|\boldsymbol{C_i})*P(\boldsymbol{C_i})) / P(D) \quad (2.2)$$

Naïve Bayes Classifier atau bisa disebut sebagai *Multinomial Naïve Bayes* merupakan model penyederhanaan dari Metode *Bayes* yang cocok dalam pengklasifikasian teks atau dokumen. Persamaannya adalah:

$$\boldsymbol{V_{MAP}} = \arg\max P(\boldsymbol{v_j} \mid \boldsymbol{a_1}, \boldsymbol{a_2}, \ldots\ldots \boldsymbol{a_n}) \quad (2.3)$$

Menurut persamaan (2.3), maka persamaan (2.1) dapat ditulis:

$$\boldsymbol{V_{MAP}} = \underset{\boldsymbol{v_j} \in \boldsymbol{V}}{\boldsymbol{argmax}} \frac{\boldsymbol{P(a_1, a_{2,\ldots}, a_n \vee v_j) P(v_j)}}{\boldsymbol{P(a_1, a_{2,\ldots}, a_n)}} \quad (2.4)$$

P($a_1, a_{2,\ldots}, a_n$) konstan, sehingga dapat dihilangkan menjadi

$$V_{MAP} = \underset{v_j \in V}{argmax} P(a_1, a_{2,\ldots}, a_n \vee v_j) P(v_j) \quad (2.5)$$

Karena P(a_1 , a_2 ,..... a_n | v_j) sulit untuk dihitung, maka akan diasumsikan bahwa setiap kata pada dokumen tidak mempunyai keterkaitan.

$$V_{MAP} = \underset{v_j \in V}{argmax} P(v_j) \prod P(a_i \vee v_j) \quad (2.6)$$

Keterangan :

$$P(v_j) = \frac{|docs_j|}{|Contoh|} \quad (2.7)$$

$$P(w_k \vee v_j) = \frac{n_k + 1}{n + |kosakata|} \quad (2.8)$$

Di mana untuk :

- P(vj) : Probabilitas setiap dokumen terhadap sekumpulan dokumen.
- P(wk|vj) : Probabilitas kemunculan kata wk pada suatu dokumen dengan kategori class vj.
- | docs | : frekuensi dokumen pada setiap kategori.
- | Contoh | : jumlah dokumen yang ada.
- Nk : frekuensi kata ke-k pada setiap kategori.
- kosakata : jumlah kata pada dokumen test.

Pada persamaan (2.8) terdapat suatu penambahan 1 pada pembilang, hal ini dilakukan untuk mengantisipasi jika terdapat suatu kata pada dokumen uji yang tidak ada pada setiap dokumen data training.

2.7. Pengujian Klasifikasi

Sebuah proses klasifikasi memerlukan proses pengujian mengenai hasil dari klasfikasi. Hal itu perlu dilakukan untuk mendapatkan akurasi dari perhitungan yang dilakukan. Proses pengujian klasifikasi menggunakan *matriks confusion* dimana penjelasan *matriks consfusion* seperti pada tabel berikut (Prasetyo Eko, 2012).

Tabel 2. *11* Matriks confusion

Keterangan	Relevan	Tidak Relevan
Terambil	True Positif (TP)	False Positif (FP)
Tidak Terambil	False Negatif (FN)	True Negatif (TN)

Dimana,

True Positif (TP) = teridentifikasi secara benar

False Positif (FP) = teridentifikasi secara salah

False Negatif = tertolak secara benar

True Negatif = tertolak secara salah

Rumus untuk menghitung akurasi seperti pada rumus 2.9 sebagai berikut.

$$Akurasi = \frac{Jumlahdatayangdiprediksibenar}{jumlahprediksiyangdilakukan} \quad (2.9)$$

$$Accurasi : \frac{TP+TN}{TP+FP+FN+TN}$$

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

$$Precision : \frac{TP+TN}{TP+FP} \quad (2.11)$$

$$F - Score : \frac{2\ x\ (Precision\ x\ Recall)}{Precision+Recall} \quad (2.12)$$

Sebuah algoritma klasifikasi berusaha untuk membentuk model yang mempunyai nilai akurasi yang tinggi. Umumnya model yang dibangun dapat memprediksi dengan benar pada semua data yang menjadi data latihnya, tetapi ketika model berhadapan dengan data uji barulah kinerja model dari sebuah algoritma klasifikasi ditentukan.

2.8. Normalisasi data

Nilai-nilai atribut data yang berbeda rentangnya seringkali perlu dilaukan normalisasi atau disandarisasikan agar proses data mining tidak menjadi bisa. Normalisasi data dapat dilakukan ke dalam rentang angka yang kecil seperti rentang [0, 1] maksudnya nilai hasil normalisasi berada diantara angka 0 (nol) sampai satu (1). Sehingga semua atribut akan memiliki bobot yang sama (Prastyo Eko, 2014).

$$X_{ik} = \frac{X_{ik} - min\ (X_k)}{max(X_k) - min(X_k)} \quad (2.13)$$

Dimana

X_{ik} = nilai asl

$Min(X_k)$ = nilai minimal (terkecil) dari kelompok k

$Max(X_k)$ = nilai maksimal (terbesar) dari kelompok k

2.9. *Technique For Others Preference by Similarity to Ideal Solution* (TOPSIS)

TOPSIS diperkenalkan pertama kali oleh Yoon dan Hwang pada tahun 1981 untuk digunakan sebagai salah satu metode dalam memecahkan masalah multikriteria (Sachdeva, 2009). Metode ini merupakan salah satu metode yang banyak digunakan untuk menyelesaikan pengambilan keputusan secara praktis. TOPSIS merupakan alternatif terbaik yang memiliki jarak terpendek dari solusi ideal positif dan jarak terjauh dari solusi ideal negatif (Hwang & Yoon, 1981 dalam Kusumadewi, 2006). dari sudut pandang geometris dengan menggunakan jarak Euclidean alternatif terbaik dari sejumlah alternatif berdasarkan beberapa kriteria tertentu. Semakin banyaknya faktor yang harus dipertimbangkan dalam proses pengambilan keputusan, maka semakin relatif sulit juga untuk mengambil keputusan terhadap suatu permasalahan. Metode ini banyak digunakan untuk menyelesaikan pengambilan keputusan secara praktis. Hal ini disebabkan konsepnya sederhana dan mudah dipahami, komputasinya efisien, dan memiliki kemampuan mengukur kinerja relatif dari alternatif-alternatif keputusan. Dengan ide dasarnya adalah bahwa alternatif yang dipilih memiliki jarak terdekat dengan solusi ideal positif dan memiliki jarak terjauh dari solusi ideal negatif.

Prosedur pehitungan dengan menggunakan metode TOPSIS :

1. Membuat matriks keputusan yang ternormalisasi.

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i}^{m} = 1^{x_{ij}^2}}} ; \textbf{dengan i} = 1, 2, \ldots\ldots, \textbf{m}; \textbf{dan j} = 1, 2, \ldots, \textbf{n}. \quad (2.14)$$

Keterangan

r_{ij} = matriks keputusan yang ternormalisasi

x_{ij} = Kriteria ke-i

2. Menentukan bobot prefrensi atau tingkat kepentingan pada kriteria

$$\textbf{Preferensi kategori} = \frac{\textbf{Jumlah data pada kategori}}{\textbf{jumlah semua data}} \quad (2.15)$$

3. Membuat matriks keputusan yang ternormalisasi terbobot

$$\boldsymbol{y_{ij} = w_i r_{ij}; \text{dengan i} = 1, 2, \dots, \text{m}; \text{dan j} = 1, 2, \dots, n.} \quad (2.16)$$

Keterangan

y_{ij} = matriks keputusan yang ternormalisasi terbobot

r_{ij} = matriks keputusan yang ternormalisasi

W_i = Bobot Kriteria ke-i

4. Menentukan matriks solusi ideal positif (A+) dan matriks solusi ideal negatif (A-) berdasarkan rating bobot ternormalisasi yij

$$\mathbf{A}^+ = (y_1^+, y_2^+, \dots, y_n^+);$$

$$A^- = (y_1^-, y_2^-, \dots, y_n^-);$$

$$y_j^+ = \begin{cases} max_i\ y_{ij} \text{ ; jika j adalah atribut keuntungan (benefit)} \\ min_i\ y_{ij} \text{ ; jika j adalah atribut biaya (cost)} \end{cases}$$

$$y_j^- = \begin{cases} min_i\ y_{ij} \text{ ; jika j adalah atribut biaya (cost)} \\ max_i\ y_{ij} \text{ ; jika j adalah atribut keuntungan (benefit)} \end{cases} \quad (2.17)$$

Keterangan :

A^+ = matriks solusi ideal positif

A^- = matriks solusi ideal positif

5. Menentukan jarak antara nilai setiap alternatif dengan matriks solusi ideal positif dan matriks solusi ideal negative

$$S_i^+ = \sqrt{\sum_j^n = 1 \left(y_i^+ - y_{ji}\right)^2} \text{ ; i} = 1, 2, \dots, \text{m}$$

$$S_i^- = \sqrt{\sum_j^n = 1 \left(y_{ji} - y_i^-\right)^2} \text{ ; i} = 1, 2, \dots, \text{m} \quad (2.18)$$

Keterangan :

S_i^+ = Jarak alternatif terhadap solusi ideal positif

S_i^- = Jarak alternatif terhadap solusi ideal negatif

6. Menentukan nilai preferensi untuk setiap alternatif (Vi)

   $$Vi = \frac{S_i^-}{S_i^- + S_i^+} \quad (2.19)$$

   Keterangan :

   Vi = Prefrensi Nilai

7. Menentukan ranking alternati, Nilai Vi yang terbesar menunjukan bahwa alternatif Ai lebih dipilih.

2.10. Penelitian Sebelumnya

Penulis mengkaji dari hasil–hasil penelitian yang memiliki kesamaan topik dengan yang sedang diteliti oleh penulis. Adapun beberapa kajian yang berhubungan dengan topik yang sedang diteliti, antara lain:

1. Jenepte Wisudawati, *"Klarifikasi Sentimen pada Movie Review dengan Metode Naive Bayes"*. Tahun 2017, Universitas Telkom. Dalam penelitian ini dilakukan penerapan metode *Naive bayes* dengan *negation handling* berdasarkan *punctuation, preprocessing dan TF-IDF* diperoleh nilai akurasi terbesar 85,16%.
2. Razzaq et al, 2014 "*Analisis sentimen pada twitter mengenai opini masyarkat terhdap pemilihan presiden Pakistan*" Pembahasan pada penelitian ini dimana sentimen tersebut dapat dijadikan sebagai acuan untuk prediksi hasil pemilu. Opini di klasifikasikan kedalam 3 kelas yaitu opini positif, negatif, dan netral. Metode *support vector machine* mendapatkan akurasi sebesar 70%.

UNIVERSITAS MUHAMMADIYAH
GRESIK