BAB II

LANDASAN TEORI

2.1. *Text Mining*

Text mining adalah istilah umum yang menggambarkan berbagai teknologi yang dapat menganalisis dan memproses data teks yang bersifat semi terstruktur (*semistructured*) dan tidak terstruktur (*unstructured*) (Miner, et al., 2012). Text mining memiliki tujuan dan menggunakan proses yang sama dengan data mining, tapi dengan tipe data yang berbeda. Data yang digunakan pada *text mining* adalah data yang tidak atau semi-terstruktur.

2.1.1. *Preprocessing*

Sebagai tahap awal dalam *text mining*, dengan 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. Tahap *preprocessing* terdiri dari beberapa proses yang akan dibahas satu per satu secara detail, antara lain (Zulfa & Winarko, 2017):

1. ***Case Folding***

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

Tabel 2. 1 Contoh Proses *Case Folding*

Input	Jokowi Memberikan Sepeda Kepada Mahasiswa
Output	Jokowi memberikan sepeda kepada mahasiswa

2. Penghapusan Tanda Baca

Penghapusan tanda baca berfungsi untuk menghapus karakter khusus dalam komentar (seperti: koma (,), titik(.), tanda tanya (?), tanda seru (!) dan sebagainya), angka numerik (0 - 9), dan karakter lainnya(seperti: $, %, *, dan sebagainya).

Tabel 2. 2 Contoh Proses Penghapusan Simbol-Simbol

Input	#Jokowi melakukan!!! pembangunan @Jakarta
Output	Jokowi melakukan pembangunan Jakarta

3. Tokenisasi

Tahap tokenisasi yaitu tahap pemotongan string inputan berdasarkan kata yang menyusunnya. Pada dasarnya proses tokenisasi adalah pemenggalan kalimat menjadi kata.

Tabel 2. 3 Contoh Proses Tokenisasi

Input	Jokowi melakukan pembangunan di kota Jakarta
Output	Jokowi, melakukan, pembangunan, di, kota, Jakarta

4. *Stopword Removal*

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

Tabel 2. 4 Contoh Proses *Stopword Removal*

Input	Jokowi melakukan pembangunan di kota Jakarta
Output	Jokowi melakukan pembangunan kota Jakarta

5. *Stemming*

Stemming merupakan proses pengubahan kata menjadi kata dasar yang sesuai dengan aturan bahasa Indonesia. Proses *stemming* menggunakan bantuan dari *RapidMiner.*

Tabel 2. 5 Contoh Proses *Stemming*

Input	Jokowi membuat program kesehatan terbaru
Output	Jokowi buat program sehat baru

2.2. *Term Frequency-Inverse Document Frequency*

TF-IDF adalah sebuah metode yang merupakan integrasi antar *term frequency* (*TF*), dan *inverse document frequency* (*IDF*). *Term Frequency* dihitung menggunakan persamaan (Allahyari, et al., 2017) dengan *term frequency* ke-i adalah frekuensi kemunculan term ke-i dalam dokumen ke-j. *Inverse Document Frequency* (*IDF*) adalah logaritma dari rasio jumlah seluruh dokumen dalam korpus dengan jumlah dokumen yang memiliki term yang dimaksud seperti yang dituliskan secara matematis pada persamaan (Romadhony, et al., 2017).

Sedangkan menurut Qaiser & Ali (2018) *TF-IDF* adalah statistik numerik yang menunjukkan relevansi dari kata kunci untuk dokumen tertentu atau dapat dikatakan bahwa, kata kunci akan didapatkan, dengan menggunakan beberapa dokumen spesifik yang dapat diidentifikasi atau dikategorikan.

Tahapan pembobotan dengan *TF-IDF* adalah:

1. Hitung *term frequency* $tf_{t,d}$

$$TF = \begin{cases} 1 + log_{10}\left(f_{t,d}\right), f_{t,d} > 0 \\ 0, f_{t,d} = 0 \end{cases}$$

Rumus 2. 1 *Term Frequency*

2. Hitung *weighting term frequency* (W_{tf})

$$W_{tft,d} = \begin{cases} 1 + log_{10} tf_{t,d}\,, if\ tf_{t,d} > 0 \\ 0, \qquad\qquad otherwise \end{cases}$$

Rumus 2. 2 *Weighting Term Frequency*

3. Hitung bobot *inverse document frequency (idf)*

$$idf_t = log_{10} \frac{N}{df_t}$$

Rumus 2. 3 *Inverse Document Frequency*

4. Hitung nilai bobot *TF-IDF*

$$W_{t,d} = W_{tf_{t,d}}\, x\, idf_t \qquad (3)$$

Rumus 2. 4 *TF-IDF*

Keterangan:

$tf_{t,d}$ = frekuensi term

$W_{tf_{t,d}}$ = bobot frekuensi term

df = jumlah frekuensi dokumen yang mengandung term

N = jumlah total dokumen

$W_{t,d}$ = bobot *TF-IDF*

2.3. *Confusion Matrix*

Confusion matrix dapat digunakan untuk mengevaluasi kualitas dari *classifier* (Han, Kamber, & Pei, 2012). *Confusion matrix* dapat diartikan sebagai suatu alat yang memiliki fungsi untuk melakukan analisis apakah *classifier* tersebut baik dalam mengenali kelas yang berbeda. Nilai dari *true positive* dan *true negative* memberikan informasi ketika *classifier* melakukan klasifikasi data bernilai benar

sedangkan *false positive* dan *false negative* memberikan informasi ketika *classifier* salah dalam melakukan klasifikasi data (Fibrianda & Bhawiyuga, 2018).

2.4. *K-Fold Cross Validation*

K-Fold Cross Validation merupakan salah satu cara untuk mengukur kualitas dari *classifier* dengan membagi data ke dalam *fold*. Metode *K-Fold Cross Validation* mempartisi himpunan data D secara acak menjadi *k-fold*(sub himpunan) yang saling bebas: $f1, f2, \dots, fk$, sehingga masing-masing *fold* berisi1/k bagian data (Suyanto, 2017). Misalnya, dengan k = 5, maka himpunan data $D1$ berisi empat *fold*, yaitu $f2$, $f3$, $f4$, dan $f5$ untuk data latih serta satu *fold* $f1$ untuk data uji. Himpunan data $D2$ berisi empat *fold*, yaitu $f1$, $f3$, $f4$, dan $f5$ untuk data latih serta satu *fold* $f2$ untuk data uji. Demikian seterusnya untuk himpunan data $D3$,$D4$,$D5$ (Suyanto, 2017).

2.5. Analisis Sentimen

Analisis sentimen adalah sebuah pemrosesan bahasa alami (PBA) dengan tugas ekstraksi informasi yang mengidentifikasi pandangan pengguna atau pendapat yang dijelaskan dalam bentuk positif, negatif, atau komentar netral dan kutipan berdasarkan text (Sekharan & Chandrakala, 2012). Analisis sentimen, juga disebut sebagai *opinion minining,* merupakan salah satu bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap orang, dan emosi terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya (Liu, 2012).

2.6. Media Sosial

Media sosial adalah media yang digunakan oleh individu agar menjadi sosial, atau menjadi sosial secara daring dengan cara berbagi isi, berita, foto dan lain-lain dengan orang lain (Taprial & Kanwar, 2012).

Media sosial adalah sebuah istilah yang menggambarkan bermacam-macam teknologi yang digunakan untuk mengikat orang-orang ke dalam suatu kolaborasi, saling bertukar informasi, dan berinteraksi melalui isi pesan yang berbasis web. Dikarenakan internet selalu mengalami perkembangan, maka berbagai macam teknologi dan fitur yang tersedia bagi pengguna pun selalu mengalami perubahan. Hal ini menjadikan media sosial lebih *hyper* dibandingkan sebuah referensi khusus terhadap berbagai penggunaan atau rancangan (Cross, 2013).

2.7 *Twitter*

Twitter adalah sebuah layanan media sosial yang memungkinkan penggunanya untuk menulis maksimal 140 karakter, yang dikenal sebagai *tweet*. *Twitter* didirikan oleh Jack Dorsey pada tahun 2006. Dilansir dari CNN Indonesia, terhitung 21 Maret 2016, *Twitter* genap memasuki usianya yang ke-10. Media sosial ini secara global memiliki sekitar 332 juta pengguna bulanan, dengan 500 juta kicauan dikirim setiap hari dan 200 miliar kicauan dalam setahun. Fitur yang terdapat pada *Twitter*, antara lain:

1. **Halaman Utama (*Home*)**

   Halaman utama merupakan kumpulan *tweets* yang di pos oleh pengguna baik informasi, berita, gambar bahkan *check-in* di suatu wilayah atau tempat dengan waktu bersamaan.

2. **Profil (*Profile*)**

   Profil merupakan kumpulan *tweet* yang pernah dibuat pengguna dan informasi mengenai data diri.

3. ***Following***

   Following merupakan akun seseorang yang mengikuti akun *Twitter* lain. *Tweet* yang di pos oleh *following* akan muncul di halaman utama.

4. ***Followers***

   Followers merupakan pengguna yang ingin berteman dan mengetahui aktivitas *tweet* seseorang. *Tweet* yang di pos oleh *followers* akan muncul di halaman utama.

5. ***Mentions***

   Konten ini merupaka balasan dari percakapan agar pengguna dapat langsung menandai orang yang akan diajak bicara.

6. ***Favorite***

   Tweets ditandai sebagai favorit agar tidak hilang dan mudah untuk dicari untuk menemukan *tweet* sebelumnya.

7. **Pesan Langsung(*Direct Message*)**

   Direct message merupakan media pengiriman pesan secara privasi antar pengguna.

8. ***Hashtag***

   Hashtag ditandai dengan "#" berfungsi untuk mencari topik yang sejenis.

9. ***List***

Pengguna *Twitter* dapat mengelompokkan *followers* ke dalam satu grup sehingga memudahkan untuk dapat melihat informasi secara keseluruhan.

10. *Trending Topic*

Topik yang sedang ramai dibicarakan oleh banyak pengguna dalam suatu waktu bersamaan.

2.8. *Support Vector Machine (SVM)*

Support Vector Machine (SVM) adalah sebuah algoritma klasifikasi yang dapat dipakai *data mining* untuk klasifikasi dan analisis regresi. Meskipun regresi dapat digunakan, sebagian besar *SVM* digunakan untuk tujuan klasifikasi (Azad et al., 2020). Berdasarkan model diskriminatif, bab ini membuktikan bagaimana *SVM* dapat menghasilkan performa yang baik pada *text classification*. Ini membuat *SVM* menjadi metode pembelajaran pertama dengan justifikasi teori untuk pengunaannya pada *text classification* (Joachims, 2012).

2.9. *Naïve Bayes Classifier (NBC)*

Naïve Bayes classifier (NBC) merupakan salah satu metode *machine learning* yang memanfaatkan perhitungan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. *Naïve bayes classifier* adalah sebuah metode sederhana yang dikembangkan berdasarkan aturan bayes, dengan melihat kondisi-kondisi yang ada dan peluang setiap kondisinya (Fanissa, Adinugroho, & Fauzi, 2018).

Naïve Bayes Clasifier atau disebut juga dengan *Bayesian Classification* merupakan metode klasifikasi statistik yang didasarkan pada teorema bayes yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas (Muhammad, 2017).

Bayesian Classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam *database* yang besar. Bentuk umum teorema bayes adalah sebagai berikut:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \qquad (1)$$

Dimana :

X = Data dengan kelas yang belum diketahui

H = Hipotesa data X merupakan suatu kelas spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (*posterior probability*)

P(H) = Probabilitas hipotesis H (*prior probability*)

Peluang bersyarat atribut kategorikal dinyatakan dalam bentuk sebagai berikut:

$$P(A_i|C_j) = \frac{|A_{ij}|}{N_{Cj}} \qquad (2)$$

Dimana $|Aij|$ adalah jumlah contoh pelatihan dari kelas Ai yang menerima nilai. Jika hasilnya adalah nol, maka menggunakan pendekatan berikut:

$$P(A_i|C_j) = \frac{n_c + n_{equiv}\,p}{n + n_{equiv}} \qquad (3)$$

Dimana n adalah total dari jumlah hasil dari kelas. nC adalah jumlah contoh pelatihan dari kelas Ai yang menerima nilai Cj. $nequiv$ adalah nilai konstan dari

ukuran sampel yang ekuivalen. *P* adalah peluang estimasi prior, *P* = 1/*k* dimana *k* adalah jumlah kelas dalam variabel target.

Peluang bersyarat atribut kontinu dinyatakan dalam bentuk berikut:

$$P(A_i|C_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}}\, exp^{\left[\frac{(A_i-\mu_{ij})^2}{2(\sigma_{ij})^2}\right]} \quad (4)$$

Rumus 2. 5 ***Naïve Bayes Classifier***

Parameter μij dapat diestimasi berdasarkan sampel *mean* Ai untuk seluruh hasil pelatihan yang dimiliki kelas. Dengan cara sama, (σij) 2 dapat diestimasi dari sampel varian ($s2$) hasil pelatihan tersebut.

2.10. *RapidMiner*

RapidMiner merupakan perangakat lunak yang bersifat terbuka (*open source*). *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining,* dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *RapidMiner* memiliki kurang lebih 500 operator *data mining*, termasuk operator untuk *input*, *output*, *data preprocessing,* dan *visualization*. *RapidMiner* merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin *data mining* yang dapat diintegrasikan pada produknya sendiri (Aprilla et al., 2013).

2.11 *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor (KNN)* adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasi sebelumya. Termasuk dalam *supervised learning*, dimana hasil *query instance* yang

baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam *KNN*.

2.12. Penelitian Terdahulu

Tabel 2. 6 Penelitian Terdahulu

Nama Peneliti	Judul Penelitian	Nama Jurnal	Hasil Penelitian	Hasil yang Diambil
Agnes Rossi Trisna Lestari, Rizal Setya Perdana, M. Ali Fauzi.	Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen *Twitter* Berbahasa Indonesia Menggunakan *Naïve Bayes* dan Pembobotan Emoji	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 1, No. 12, 2017.	Pembobotan non tekstual mempengaruhi hasil dari klasifikasi sentimen, dan penggabungan pembobotan dapat meningkatkan hasil akurasi.	Penggabungan pembobotan tekstual dan non tekstual.
Achmad Bayhaqy, Sfenrianto Sfenrianto, Kaman Nainggolan, Emil R Kaburuan.	*Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes*	2018 *International Conference on Orange Technologies, ICOT 2018*	Klasifikasi analisis sentimen dari *Twitter* mengenai tanggapan *customer* terhadap *marketplace online*	Klasifikasi analisis sentimen menggunakan *RapidMiner*
Winda Estu Nurjanah, Rizal Setya Perdana, Mochammad Ali Fauzi.	Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial *Twitter* menggunakan	Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 1, No. 12, 2017	Tingkat akurasi untuk analisis sentimen meningkat ketika menggunakan metode *K-Nearest Neighbor* dan penggabungan	Penggabungan pembobotan tekstual dan non tekstual serta menggunakan metode tambahan.

	Metode *K-Nearest Neighbor* dan Pembobotan Jumlah *Retweet*		fitur pembobotan tekstual dan non-tekstual.	

Dari tabel 2.6 penelitian terdahulu (Rossi et al., 2017) melakukan analisis sentimen dengan penggabungan pembobotan tekstual dan non-tekstual dapat mempengaruhi hasil akurasi. Hasil akurasi tekstual 68,52% untuk pembobotan tekstual, 75,93% untuk pembobotan non-tekstual, dan 74,81% setelah melakukan penggabungan pembobotan tekstual dan non tekstual. Hasil penelitian terdahulu (Nurjanah et al., 2017) akurasi untuk pembobotan tekstual sebesar 82,50%, untuk pembobotan non-tekstual sebesar 60%, setelah melakukan penggabungan pembobotan akurasi menjadi 83,33%. Pada penelitian terdahulu (Bayhaqy et al., 2018) melakukan sentimen analisis dan membandingkan tiga metode klasifikasi yaitu *Decision Tree, K-NN,* dan *Naïve Bayes* untuk mencari akurasi terbaik menggunakan aplikasi *Rapidminer*. Hasil penelitian *Naïve Bayes* mendapat akurasi paling tinggi sebesar 77%.