

BAB II
TINJAUAN PUSTAKA

2.1 Dasar Teori

2.1.1 Text Mining

Text mining dapat didefinisikan secara luas sebagai suatu proses dimana pengguna berinteraksi dengan koleksi dokumen menggunakan suatu *tool* analisis. Dalam kaitannya dengan *data mining*, *text mining* merupakan suatu proses untuk mengekstrak informasi yang berguna dari suatu sumber data melalui identifikasi dan eksplorasi pola tertentu (Feldman & Sanger, 2006).

Tujuan dari *text mining* yaitu untuk memproses informasi tekstual yang tidak terstruktur, mengekstrak indeks numerik yang bermakna dari teks, dan kemudian membuat informasi yang terkandung di dalam teks dapat diakses menggunakan berbagai algoritma *data mining*.

Menurut Miner (Miner, et al., 2012) proses *text mining* dapat dikelompokkan menjadi tujuh area, yaitu :

1. Pencarian dan perolehan informasi (*search and information retrieval*), yaitu proses yang meliputi indeksing, pencarian, dan perolehan dokumen dari suatu database menggunakan kata kunci *(keyword).*
2. Pengelompokan dokumen (*document clustering),* yaitu proses pengelompokan suatu dokumen berdasarkan pada kemiripan kata antar dokumen menggunakan algoritma *clustering*.
3. Klasifikasi dokumen *(document classification),* merupakan suatu proses pengelompokan dan pengkategorian suatu dokumen berdasarkan model terlatih yang sudah memiliki label sebelumnya.
4. *Web mining,* yaitu proses penggalian informasi yang muncul di internet dalam jumlah yang besar dengan skala fokus yang spesifik.
5. Ekstraksi informasi *(information extraction),* yaitu proses yang bertujuan untuk mengekstrak suatu informasi dari data yang sifatnya semi-struktur ataupun tidak terstruktur menjadi data yang terstruktur.

6. *Natural Language Processing* (NLP), proses *text mining* yang bertujuan untuk membuat suatu program yang memiliki kemampuan untuk memahami bahasa manusia.
7. Ekstraksi konsep *(concept extraction)*, yaitu proses untuk mengelompokkan kata yang memiliki kemiripan secara semantik.

2.1.2 Text Preprocessing

Text preprocessing merupakan suatu proses pengubahan bentuk data tekstual yang belum terstruktur menjadi data yang terstruktur dan disimpan dalam basis data (Langgeni, Baizal, & A.W., 2010).

Tahapan *text preprocessing* pada penelitian ini terdiri dari beberapa langkah yaitu :

1. *Case Folding*

   Proses *case folding* merupakan proses untuk menghilangkan semua karakter selain huruf (seperti angka dan tanda baca) dan mengubah semua huruf menjadi huruf kecil.
2. *Tokenization*

   Tokenization merupakan proses pemotongan kalimat berdasarkan tiap kata yang menyusunnya.
3. *Stemming*

   Stemming merupakan proses pemotongan imbuhan atau pengembalian kata berimbuhan menjadi kata dasar. Proses *stemming* dalam penelitian ini menggunakan Algoritma Nazief Adriani yang diambil dari library Sastrawi (Librian, 2004).
4. *Filtering*

   Filtering atau disebut juga *stopword removal* merupakan proses untuk menghilangkan *stopwords* (kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words). Proses ini bertujuan untuk mengurangi jumlah kata.

2.1.3 Document Frequency Thresholding

Salah satu masalah yang umum ditemukan dalam proses klasifikasi maupun *clustering* dokumen adalah tingginya dimensi data, sehingga perlu dilakukan proses seleksi fitur untuk memilih beberapa fitur yang dapat digunakan untuk mewakili dokumen (Langgeni, Baizal, & A.W., 2010).

Salah satu teknik seleksi fitur yang paling sederhana namun memiliki kinerja yang cukup baik adalah *Document Frequency Thresholding* yang bersifat *class independent. Document Frequency* merupakan banyaknya jumlah dokumen yang mengandung *term* tertentu. *Term* yang jarang muncul memiliki kemungkinan besar tidak memberikan informasi spesifik. Begitupun jika *term* tersebut terlalu sering muncul pada banyak dokumen, maka dianggap bahwa *term* tersebut merupakan *term* yang umum dan tidak akan mempengaruhi kinerja pediksi secara keseluruhan (Nallaswamy, 2012).

2.1.4 Term Frequency – Inverse Document Frequency (TFIDF)

Term weighting atau pembobotan kata bertujuan untuk memberikan bobot nilai pada setiap kata. Perhitungan bobot ini memerlukan dua hal, yaitu *Term Frequency (TF)* dan *Inverse Document Frequency (IDF). Term Frequency* merupakan banyaknya jumlah kata atau *term* tertentu yang ada dalam suatu dokumen. Sementara *Inverse Document Frequency* adalah frekuensi kemunculan kata atau *term* pada seluruh dokumen. Nilai IDF berbanding terbalik dengan jumlah dokumen yang mengandung *term* tertentu. *Term* yang jarang muncul pada seluruh dokumen memiliki nilai IDF yang lebih besar dari nilai IDF *term* yang sering muncul. Jika pada setiap dokumen mengandung *term* tertentu, maka nilai IDF *term* tersebut bernilai 0. Hal ini menunjukan bahwa *term* yang muncul pada seluruh dokumen merupakan *term* yang tidak berguna untuk membedakan dokumen berdasarkan topik tertentu (Luthfiarta, Zeniarja, & Salam, 2013).

Rumus TF-IDF adalah sebagai berikut :

$$W_{dt} = tf_{dt} \; x \; idf_t = tf_{dt} \; x \; log \; (\frac{N}{df_t}) \qquad (2.1)$$

Dimana :

$W_{d,j}$ = bobot *term* ke-t terhadap dokumen d

tf_d = jumlah kemunculan *term* t dalam dokumen d

N = jumlah dokumen secara keseluruhan

df_t = jumlah dokumen yang mengandung *term* t

2.1.5 Naive Bayes Classifier

Pengelompokan dokumen dapat dilakukan dengan tiga cara yaitu, *supervised, unsupervised,* dan *semi supervised.* Salah satu kegiatan penting dalam pengelompokan atau kategorisasi teks adalah dengan pendekatan *supervised.* Kategorisasi teks sendiri saat ini memiliki berbagai cara pendekatan antara lain berbasis numeris, misalnya pendekatan *probabilistic*, *support vector machine*, *artificial neural network*, serta berbasis non numeris seperti *decision tree classification.* Dari kelompok pendekatan berbasis numeris, pendekatan berbasis *probabilistic Naive Bayes Classifier (NBC)* memiliki beberapa kelebihan antara lain, sederhana, cepat dan memiliki akurasi yang tinggi (Hamzah, 2012).

Naive Bayes merupakan metode yang berdasarkan pada teorema Bayes yang dirumuskan pada persamaan 2.1 berikut ini (Jurafsky, 2011) :

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)} \qquad (2.2)$$

Keterangan :

c = hipotesis d pada label tertentu

d = data kelas yang belum diketahui

P(c) = probabilitas *prior* dari c

P(d) = probabilitas *prior* dari d

P(c|d) = probabilitas c berdasarkan kondisi d

P(d|c) = probabilitas d pada hipotesis c

2.1.6 Multinomial

Model multinomial memperhitungkan frekuensi setiap kata yang muncul pada dokumen. Misal terdapat dokumen d dan himpunan kelas c. Untuk memperhitungkan kelas dari dokumen d, maka dapat dihitung dengan rumus :

$$P(c|\text{term dokumen } d) = P(c)\ x\ P(t_1|c)\ x\ P(t_2|c)\ x\ P(t_3|c)\ x\ \ldots\ x\ P(t_n|c) \qquad (2.3)$$

Keterangan :

P(c|term dokumen d) = Probabilitas suatu dokumen termasuk kelas *c*

P(c) = Probabilitas *prior* dari kelas c

t_n = Kata dokumen d ke-n

P(t_n|*c*) = Probabilitas kata ke-n dengan diketahui kelas *c*

Probabilitas *prior* kelas c ditentukan dengan rumus :

$$P(c) = \frac{N_c}{N} \quad (2.4)$$

Keterangan :

Nc = Jumlah kelas c pada seluruh dokumen

N = Jumlah seluruh dokumen

Probabilitas kata ke-n ditentukan dengan menggunakan teknik *laplacian smoothing* :

$$P(t_n \mid c) = \frac{count(tn,c)+1}{count(c)+|V|} \quad (2.5)$$

Keterangan :

count(t_n,*c*) = Jumlah term t_n yang ditemukan di seluruh data pelatihan dengan kategori c

count(c) = Jumlah term di seluruh data pelatihan dengan kategori c

V = Jumlah seluruh tem pada data pelatihan

Adapun rumus Multinomial yang digunakan dengan pembobotan kata TF-IDF adalah sebagai berikut :

$$P(t_n \mid c) = \frac{Wct+1}{(\sum W' \epsilon V W' ct)+B'} \quad (2.6)$$

Keterangan :

Wct = Nilai pembobotan tfidf atau W dari *term* t di kategori c

$\sum W' \in V W' ct$ = Jumlah total W dari keseluruhan *term* yang berada di kategori c.

B' = Jumlah W kata unik (nilai idf tidak dikali dengan tf) pada seluruh dokumen.

2.2 Penelitian Terkait

Beberapa penelitian sebelumnya yang terkait dengan pengelompokan dokumen teks diantaranya :

Penelitian dengan judul "Klasifikasi Teks dengan Naive Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis" yang dilakukan oleh Amir Hamzah (Hamzah, 2012). Penelitian menggunakan data 1000 dokumen berita dan 450 dokumen abstrak akademis. Hasil penelitian ini menunjukkan bahwa penggunaan kata unik dalam koleksi dokumen latih tanpa filter memberikan kinerja yang kurang optimal. Adapun akurasi maksimal yang didapatkan adalah sebesar 91% untuk klasifikasi dokumen berita dan 82% untuk dokumen akademik.

Penelitian selanjutnya adalah "Klasifikasi Emosi untuk Teks Bahasa Indonesia menggunakan Metode Naive Bayes" yang dilakukan oleh (Sumpeno & Destuardi, 2009). Penelitian ini bertujuan untuk mengklasifikasikan suatu informasi verbal ke dalam beberapa kelas emosi yaitu jijik, malu, marah, sedih, senang, dan takut menggunakan metode multinomial. Dari hasil pengujian yang dilakukan dapat disimpulkan bahwa teknik multinomial cukup baik untuk kategorisasi teks Bahasa Indonesia dengan akurasi sebesar 61,57%.

Selain itu, penelitian yang dilakukan oleh (Shaufiah, Imanudin, & Asror, 2016) dengan judul *"Android Short Messages Filtering for Bahasa Using Multinomial Naive Bayes"*. Penelitian ini bertujuan untuk melakukan klasifikasi SMS antara SMS *spam* dan *not spam* dengan menggunakan metode *Multinomial Naive Bayes* dan pembobotan *TFIDF*. Hasil akurasi yang didapatkan sangat tinggi, yaitu sebesar 94,44%.

Selanjutnya, penelitian dengan menggunakan *DF-Thresholding* sebagai *feature selection* pernah dilakukan oleh (Nallaswamy, 2012) dengan judul "*A Study on Analysis of SMS Classification Using Document Frequency Threshold"*. Fokus penelitian ini adalah klasifikasi teks berbasis SVM dengan menggunakan *document frequency thresholding* sebaga metode untuk menyeleksi fitur. Penelitian ini menggunakan teks pesan NUS SMS sebagai data set. Dari hasil penelitian ini dapat disimpulkan bahwa *document frequency thresholding* cukup sederhana dan efisien untuk menyeleksi fitur dalam pengelompokan dokumen.

2.3 Rencana Penelitian

Mengacu pada penelitian-penelitian sebelumnya, penelitian kali ini megangkat tema pengelompokan berita *online* berbasis pada klasifikasi teks dengan *Multinomial Naive Bayes* dan *Document Frequency Thresholding*, *TFIDF* serta *kombinasi DF Thresholding* dan *TFIDF* sebagai metode untuk menyeleksi fitur.

Adapun keterkaitan antara penelitian ini dengan penelitian-penelitian yang dilakukan sebelumnya dapat dilihat pada Tabel 2.1.

Tabel 2.1 Keterkaitan Penelitian

No	Judul Penelitian	Tujuan Peneltian	Pengambilan Konsep
1.	Klasifikasi Teks dengan *Naive Bayes Classifier* (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis (Hamzah, 2012).	Untuk mengkaji metode *Naive Bayes Classifier* dalam pengelompokan dokumen berita dan abstract akademis	Pengelompokan dokumen teks Bahasa Indonesia dengan *Naive Bayes Classifier*.
2.	Klasifikasi Emosi untuk Teks Bahasa Indonesia menggunakan Metode Naive Bayes (Sumpeno & Destuardi, 2009)	Untuk mengkaji metode *Naive Bayes* model *Multinomial* untuk melakukan pengelompokan informasi verbal berbahasa Indonesia	Pengelompokan dokumen teks Bahasa Indonesia dengan *Naive Bayes Classifier* model *Multinomial*.
3.	*Android Short Messages Filtering for Bahasa Using Multinomial Naive Bayes* (Shaufiah, Imanudin, & Asror, 2016)	Untuk melakukan pengelompokan SMS *spam* dan *not spam* menggunakan *Multinomial Naive Bayesi* dan pembobotan TFIDF	Pengelompokan dokumen teks Bahasa Indonesia dengan *Naive Bayes Classifier* model *Multinomial* dan pembobotan TFIDF.
4.	*A Study on Analysis of SMS Classification Using Document Frequency Threshold* (Nallaswamy, 2012).	Untuk melakukan pengelompokan teks berbasis SVM menggunakan *DF Thresholding* sebagai metode untuk menyeleksi fitur.	Penggunaan seleksi fitur dengan metode *Document Frequency Thresholding* pada pengelompokan dokumen.