BAB II

TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian terdahulu menjadi salah satu acuan sehingga dapat memperkaya teori yang yang digunakan dalam mengkaji penelitian yang dilakukan. Dari penelitian terdahulu, tidak ditemukan judul yang sama dengan penelitian ini. Namun beberapa penelitian diangkat untuk dijadikan sebagai referensi dalam memperkaya bahan kajian pada penelitian. Berikut penelitian terdahulu berupa beberapa jurnal terkait dengan penelitian ini.

Tabel 2. 1 Penelitian Terdahulu

No	Nama Peneliti	Judul	Metode	Hasil
1	Miftahul Kahfi Al Fath	Analisis Sentimen Komentar Kebijakan *Full Day School* Dari *Facebook Page* Kemendikbud Ri Menggunakan Algoritma *Naïve Bayes Classifier*	Algoritma *Naïve Bayes Classifier*	Semakin besar data latih dan seleksi fitur yang digunakan pada Algoritma *Naïve Bayes Classifier* (NBC) mempengaruhi hasil akurasi algoritma.

2	Antonius Rachmat, Yuan Lukito	Klasifikasi Sentimen Komentar Politik dari *Facebook Page* Menggunakan *Naive Bayes*	*Naive Bayes*	Algoritma *Naive Bayes* mampu mengklasifikasikan sentimen dengan tingkat akurasi rata-rata tertinggi 82%
3	Fajar Ratnawati	Implementasi Algoritma *Naive Bayes* Terhadap Analisis Sentimen Opini Film Pada *Twitter*	Algoritma *Naive Bayes*	Hasil akurasi akan semakin tinggi dan itu menandakan sistem berhasil melakukan klasifikasi dengan baik.

2.2 Teori Terkait

2.2.1 Facebook

Facebook adalah sebuah layanan jejaring sosial dan situs web yang diluncurkan pada 4 Februari 2004 yang dioperasikan dan dimiliki oleh *Facebook, Inc*. *Facebook* lahir atas usaha seorang mantan mahasiswa Harvard bernama Mark Zuckerberg. Mark Zuckerberg menciptakan *Facemash*, pendahulu *Facebook*, tanggal 28 Oktober 2003 ketika berada di Harvard sebagai mahasiswa tahun kedua. *Facemash* menarik 450 pengunjung dan 22.000 tampilan foto pada empat jam pertama mengudara[5]. *Facebook* sebagian salah satu situs jejaring sosial yang populer, mempunyai nilai tersendiri bagi para penggunanya. *Facebook* sendiri tercatat mengalami kenaikan jumlah pengguna yang pesat semenjak awal didirikan. Hanya dalam kurun waktu 8 tahun semenjak didirikan pada tahun 2004, *facebook*

mencatat 835.525.280 pengguna di penjuru dunia. Angka ini berdasar laporan dalam internet *Worlds Stats,* sebuah lembaga statistic independen dari *Miniwatss Marketing Group*.

2.2.2 Analisis Sentimen

Sentiment Analysis atau *opinion mining* merupakan salah satu cabang ilmu dari *text mining*, natural language program, dan *artificial intelligence*. Proses yang dilakukan pada *Sentiment Analysis* adalah memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini[6]. Analisis sentimen bertujuan untuk melakukan penilaian terhadap pendapat tokoh masyarakat berdasarkan data tekstual. Fenomena pertumbuhan data yang terjadi secara eksponensial menjadi tantangan baru dalam analisis sentimen. Pendekatan secara konvensional bukan lagi jawaban yang tepat untuk menentukan jenis sentimen dalam data tekstual. Mempekerjakan manusia untuk mengklasifikasikan jenis sentimen dari suatu kumpulan data tekstual yang sangat besar dan beragam tentu akan membutuhkan biaya dan waktu yang tidak sedikit.

2.2.3 Preprocessing

Pre-processing merupakan tahap awal dari *text mining* untuk mengubah data sesuai dengan format yang dibutuhkan. Proses ini dilakukan untuk menggali, mengelola dan mengatur informasi dan untuk menganalisis hubungan tekstual dari data terstruktur dan data tidak terstruktur[7].

2.2.3.1 Case Folding

Case Folding merupakan tahap awal pada *Preprocessing* yang bertujuan untuk mengubah kata menjadi huruf kecil atau *lower case*

2.2.3.2 Tokenizing

Tokenizing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Sebagai contoh karakter *whitespace*, seperti *enter*, tabulasi, spasi dianggap sebagai pemisah kata.

2.2.3.3 Filtering

Filtering adalah tahap mengambil kata-kata penting dari hasil token. Bisa menggunakan algoritma *stoplist* (membuang kata kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist/stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh stopwords adalah "yang", "dan", "di", "dari" dan bahkan emoji.

2.2.3.4 Stemming

Stemming merupakan proses untuk mencari stem (kata dasar) dari kata hasil *filtering*. Terdapat dua aturan dalam melakukan *stemming* yaitu dengan pendekatan kamus dan pendekatan aturan.

2.2.4 Pembobotan TF-IDF

TF-IDF (*Term Frequency – Inverse Document Frequency*) adalah teknik pembobotan yang sering diterapkan di berbagai permasalahan penggalian informasi. Metode TF-IDF menggabungkan dua konsep untuk perhitungan bobot, yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut [8]. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting

kata tersebut di dalam dokumen. Sedangkan frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut. Sehingga bobot hubungan kata dengan dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut menjadi rendah pada kumpulan dokumen[9]. Perhitungan untuk mencari nilai TF-IDF menggunakan rumus berikut :

$$tf_{t,d} = \frac{f_d}{maxf_{\ d}}$$

$$idf_t = log\frac{N}{df_t}$$

$$W_{t,d} = tf_{t,d}\ x\ idf_t$$

Keterangan :

N : dokumen ke-N

f_d : frekuensi term (t) pada document (d)

max f_d : frekuensi maksimal term (t) pada document (d)

df_t : jumlah dokumen yang mengandung t

idf_t : nilai IDF

$tf_{t,d}$: nilai frekuensi kemunculan t pada dokumen d(TF)

$W_{t,d}$: nilai bobot dari t dalam satu dokumen (TF-IDF)

2.2.5 Naive Bayes

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. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan *Naive Bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Data *Training*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian.

Proses *Naive Bayes* dibagi kedalam 2 proses yaitu proses *training* dan *testing*. Proses *training* digunakan untuk menghasilkan model sentimen analisis yang nantinya akan digunakan sebagai pedoman dalam klasifikasi dengan data *testing* atau data yang berbeda. Berikut adalah algoritma klasifikasi untuk proses *training* dan *testing* pada algoritma *Naive Bayes*.

a) Proses *Training*

Pada proses *Naive Bayes* menggunakan rumus sebagai berikut :

$$P(Vj) = \frac{|docs\ j|}{|contoh|}$$

$$P(Xi|Vj) = \frac{nk + 1}{n + |kosakata|}$$

Keterangan :

docs j| : jumlah dokumen pada kategori j

|contoh| : jumlah dokumen dari semua kategori

Nk : jumlah kemunculan kata xi pada kategori Vj

n : jumlah kata dalam setiap kategori

|kosakata| : jumlah semua kata dari semua kategori

Pada rumus di atas bertujuan untuk menghitung bobot atau nilai probabilitas setiap kata dalam data *training* di setiap kategori klasifikasi, kemudian setiap nilai probabilitas kata tersebut digunakan dalam proses testing.

b) Proses *Testing*

Pada proses *testing* dalam algoritma *Naive Bayes* menggunakan rumus sebagai berikut:

$$Vmap = \ argmax \prod_{i=I}^{n} \quad P(Xi, |Vj)P(Vj)$$

Keterangan :

Kategori komentar j = 1, 2, 3,….n. Dimana dalam penulisan ini j_1 kategori komentar sentimen positif, j_2 = kategori komentar sentimen negatif dan

j_3= kategori komentar sentimen netral

Vmap : Semua Kategori yang diujikan V

P(xi,Vj) : Probabilitas Xi pada kategori Vj

P(Vj) : Probabilitas dari Vj

Dengan rumus di atas, setiap nilai dari kata pada dokumen testing akan dihitung berdasarkan nilai probabilitas setiap kata yang dihasilkan dari proses *training*. Perhitungan dengan rumus di atas, dilakukan untuk setiap kategori klasifikasi kemudian dicari *Vmap* tertinggi.

2.2.6 Confusion Matrix

Untuk melakukan pengujian terhadap sistem, dilakukan evaluasi akurasi sistem dalam mengklasifikasikan sentimen pada dataset dengan menggunakan

confusion matrix[10]. Dengan *confusion matrix* dapat dianalisa seberapa baik classifier dapat mengenali record dari kelas-kelas yang berbeda.

Tabel 2.2 Confusion Matrix

		Prediksi		
		Positif	Negatif	Netral
Aktual	Positif	TPos	FPosNeg	FPosNet
	Negatif	FNegPos	TNeg	FNegNet
	Netral	FNetPos	FNetNeg	TNet

Tidak jauh berbeda dengan yang berdimensi 2 x 2, multiclass confusion matrix juga memiliki elemen TP (*True Positive*), FN (*False Negative*), FP (*False Positive*), dan TN (*True Negative*). Berikut adalah ketentuan dalam menetapkan nilai elemen tersebut:

a. TP (*True Positive*) merupakan banyaknya data yang kelas aktualnya sama dengan kelas prediksinya.
b. FN (*False Negative*) merupakan total dari seluruh baris yang ditunjuk kecuali TP yang dicari.
c. FP (*False Positive*) merupakan total dari seluruh kolom yang ditunjuk kecuali TP yang dicari.
d. TN (*True Negative*) merupakan total dari seluruh kolom dan baris selain yang ditunjuk.

Accuracy adalah sebuah metode pengujian berdasarkan pada tingkat kedekatan antara nilai aktual dengan nilai prediksi. Akurasi hasil prediksi dapat diketahui dengan mengetahui jumlah data yang diklasifikasikan dengan benar. Berikut ini adalah persamaan dari akurasi :

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \textbf{*100\%}$$