Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

Perpustakaan Univeristas Multimedia Nusantara

BAB II
LANDASAN TEORI

2.1 Analisis Sentimen

Analisis sentimen adalah salah satu pembelajaran *machine learning* (komputasi mesin) yang mempelajari mengenai emosi, pendapat, penilaian atau suatu pandangan dari kumpulan teks untuk mengidentifikasi suatu karakteristik dari kumpulan kata. Proses dari analisis sentimen ini adalah mengklasifikasi dokumen berbentuk teks ke dalam suatu kelas positif atau negatif [8].

2.2 *Text Mining*

Text mining atau dalam Bahasa Indonesia penambangan teks adalah suatu proses penambangan teks yang dilakukan oleh sistem komputer untuk mendapatkan sesuatu informasi yang kemudian akan diproses informasi tersebut menjadi suatu data yang berguna. Penambangan data dapat dilakukan dengan menggunakan aplikasi *Rapidminer Studio* [9].

Text mining bertujuan untuk mendapatkan suatu informasi yang berguna dari sekumpulan data atau dokumen yang diolah. Hasil dari Text mining ini adalah kumpulan data berupa text yang tidak terstruktur dan tidak memiliki format yang akan digunakan untuk membuat kategori teks dan pengelompokan teks [9].

2.3 Media Sosial *Twitter*

Media sosial muncul sejak tahun 1978 merupakan salah satu platform online yang paling banyak digunakan masyarakat di dunia maupun Indonesia. Sosial media merupakan tempat dimana masyarakat dapat mengekspresikan pendapat,

opini, berita dan hal lainnya. Masyakat dapat dengan mudah mengakses sosial media serta menjangkaunya tanpa batasan waktu. Hal ini didukung oleh mudahnya proses pendaftaran, mendapatkan, dan mengakses sosial media secara daring sehingga tidak heran jika pengguna sosial media semakin hari akan semakin bertumbuh [10].

Gambar 2. 1 Logo *Twitter*

Gambar 2.1 adalah logo *Twitter*. *Twitter* muncul sejak tahun 21 Maret 2006 dikenal sebagai jejaring sosial yang penggunanya dapat membuat atau memposting seperti blog dengan nama lain *tweet*. *Twitter* adalah media sosial yang memungkinkan penggunanya untuk mengirimkan pesan secara *realtime* seperti perasaan dan opininya mengenai banyak isu, permasalahan, dan hal-hal lainnya. [11]

2.4 Toko Online

Toko online merupakan pasar yang dapat dijangkau dengan mudah, hanya dengan menggunakan ponsel pintar dan koneksi internet, maka dengan mudahnya

kita dapat terhubung dengan pasar. Tidak perlu repot-repot lagi ke pasar untuk membeli perlengkapan dan kebutuhan kita. Cukup dengan pilih produk kemudian bayar, lalu pesanan dating diantar oleh kurir. Transaksi jual beli yang dulunya konvensional, kini sudah berubah ke tingkat yang lebih modern yang lebih dikenal dengan *e-commerce*. [12]

2.5 *Text Preprocessing*

Tahapan text preprocessing merupakan lanjutan dari tahap penambangan data. Hasil penambangan data tentunya belum rapi dan tidak terstruktur. Maka dari itu dibutuhkan tahap text preprocessing dimana tahap ini akan mengolah informasi yang kotor tadi, akan melalui serangkaian proses seperti berikut untuk menghasilkan data yang dapat digunakan untuk penelitian:

2.5.1 *Data Cleansing*

Pada tahap ini, dilakukan penghapusan karakter-karakter selain yang sudah ditentukan seperti huruf atau karakter di luar dari daftar alfabet a sampai dengan z termasuk tanda baca pada data *tweet*. Contohnya seperti “sHopee banyak diskonnya enggaa,,??” menjadi “shopee banyak diskonnya enggaa”.

2.5.2 *Data Labelling*

Tahap *labelling* data akan dilakukan secara manual. Data *tweet* yang sudah terkumpul, *tweet* tersebut akan diberikan label yang terbagi atas empat kelas yaitu *True Positive, True Negative, Fake Positive, atau Fake Negative.*

2.5.3 *Case Folding*

Pada tahap ini, data *tweet* akan dilakukan pemrosesan berupa pengubahan seluruh teks huruf kapital atau huruf besar menjadi huruf kecil.

2.5.4 *Tokenizing*

Proses tokenizing ini akan memisahkan sebuah teks panjang berupa paragraf atau kalimat menjadi teks terpisah yang berbentuk suatu bagian-bagian kecil yang biasanya disebut token untuk dianalisa. (Yanis, 2018) Contohnya seperti berikut:

Kalimat = “shopee lagi ada banyak promo voucher nih”

menjadi = [‘shopee’, ‘lagi’, ‘ada’, ‘banyak’, ‘promo’, ‘voucher’, ‘nih’]

Pada tahap tokenizing ini juga akan dilakukan penghapusan tanda baca, karena tanda baca akan mengganggu proses perhitungan dalam algoritma yang akan diterapkan

2.5.5 TF-IDF

Pembobotan TF-IDF (*Term Frequency-Inverse Document Frequency)* merupakan salah satu proses dari teknik ekstraksi fitur dengan memberikan nilai tertentu pada setiap kata yang ada pada data latih *tweets.* Dalam mengetahui seberapa penting sebuah kata mewakili sebuah kalimat, maka dilakukan pembobotan atau perhitungan terhadap kata dalam sebuah kalimat [12].

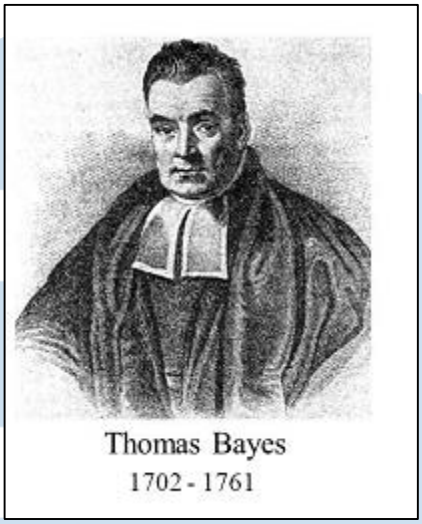
2.5.6 N-Gram character

N-Gram adalah pendekatan identifikasi dan analisis fitur populer yang sering digunakan dalam pemodelan bahasa dan pemrosesan bahasa alami. Membagi kalimat menjadi beberapa kata tertentu lalu dilihat pola huruf dalam setiap kalimat pada data. Metode N-Gram ini telah digunakan oleh beberapa penelitian lainnya yang menghasilkan penambahan nilai akurasi [13].

2.6 *Naïve Bayes Algorithm*

Algoritma *Naïve Bayes* ditemukan pada pertengahan abad ke-18 yang ditemukan oleh Reverend Thomas Bayes. Naïve bayes populer disebut sebagai metode pengelompokan teks dan pengkategorian menggunakan frekuensi kata-kata. Naive Bayes merupakan klasifikasi statistik yang memprediksi keanggotaan kelas dimana sampel yang ada akan termasuk ke dalam kelas tertentu. Algoritma ini akan digunakan pada penelitian ini untuk mengklasifikasikan data teks yang diambil dari *Twitter* kemudian teks tersebut akan diklasifikasikan menjadi kelas positif atau kelas negatif. Naive bayes merupakan metode klasifikasi yang berdasarkan pada teorema Bayes yaitu dapat memprediksi peluang dimasa yang akan datang berdasarkan data yang sudah didapat di masa lalu [9]. Gambar 2.2 adalah Thomas Bayes.

UNIVERSITAS MULTIMEDIA NUSANTARA



Gambar 2. 2 Thomas Bayes

Naïve Bayes adalah algoritma pengklasifikasian probabilistik yang menghitung sekumpulan probabilitas dengan mengkombinasikan nilai dari dataset yang diberikan lalu frekuensinya dijumlahkan. Algoritma teorema *bayes* mengasumsikan bahwa semua atribut yang ada adalah atribut independen atau tidak saling ketergantungan pada nilai yang diberikan pada variabel kelas. Kelebihan algoritma *Naïve Bayes* salah satunya adalah menghasilkan akurasi yang cukup baik dalam mengolah data besar seperti sentimen analisis. Rumus 2.2 adalah rumus *Naïve Bayes* [14].

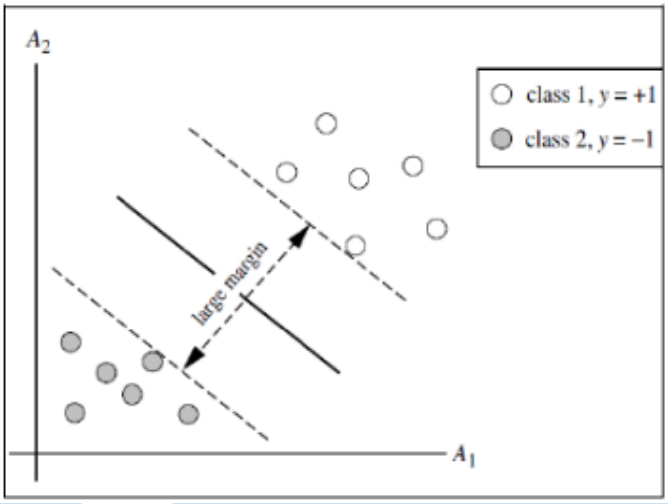
Rumus 2. 1 Rumus *Naïve Bayes*

$$fc(E) = \underset{c \in C}{argmax}\, P(c) \prod_{j=1}^{n} P(Ej|c)$$

2.7 *Support Vector Machine* (SVM)

Support Vector Machine adalah salah satu metode klasifikasi menggunakan machine learning yang memprediksi kelas dan pola dari hasil proses data training.

Klasifikasi SVM dilakukan dengan cara mencari batas yang memisahkan antara kelas positif dan kelas negative [15].



Gambar 2. 3 Pemisahan batas *Support Vector Machine*

Gambar 2.1 [16] merupakan gambaran kerja SVM yaitu menentukan batas pada kelas data. SVM menggunakan ruang hipotesis berupa fungsi linear dalam ruang berdimensi tinggi dalam sistem pengklasifikasiannya. Pada ruang yang memiliki dimensi tinggi akan dicari fungsi garis pemisah (*hyperlane*) yang dapat memaksimalkan jarak antara kelas data [17].

2.8 *Logistic Regression*

Regresi Logistik (*Logistic Regression*) adalah metode analisis yang digunakan untuk menganalisis relasi antara satu atau beberapa variabel yang hanya mempunyai variabel respon yang bebas. *Logistic Regression* memiliki dua jenis klasifikasi yaitu klasifikasi biner dimana klasifikasi terdiri dari dua kelas dan klasifikasi multinomial yang memungkinkan algoritma melakukan klasifikasi lebih dari dua kelas [18].

Rumus 2. 2 Rumus Regresi Logistik

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots.+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots.+\beta_k x_k}},$$

Rumus 2.2 adalah rumus yang digunakan algoritma Regresi Logistik. Tujuan regresi logistic adalah untuk memperkirakan kemungkinan kejadian tertentu yang terjadi didalam suatu populasi[19].

2.9 *K-Fold cross validation*

Validasi akurasi yang digunakan dalam penelitian ini adalah k-fold cross validation. Proses validasi dilakukan untuk membuktikan bahwa suatu proses atau metode dapat memberikan hasil yang konsisten dan mencapai hasil dari algoritma yang sudah diterapkan sehingga dapat menghasilkan hasil akurasi yang lebih baik [16].

2.10 *Rapidminer Studio*

Rapidminer studio adalah perangkat lunak atau *software* yang digunakan untuk pengelolaan data. Prinsip kerja yang digunakan *Rapidminer Studio* adalah dengan mengekstrak pola-pola dari data set besar kemudian dikombinasikan dengan metode statistika, kecerdasan buatan, dan *database* [20]

Gambar 2. 4 Logo *Rapidminer*

Gambar 2.4 adalah logo aplikasi *Rapiminer*. *Rapidminer* dapat melakukan perhitungan data yang sangat banyak dengan bantuan operator-operator yang disediakan dan dapat menambah operator dengan menambahkan *plugin*. Operator ini berfungsi untuk modifikasi data yang kemudian operator ini disambungkan ke operator lain dengan parameter khusus yang dapat diubah sesuai dengan kebutuhan didalam operator [20].

2.11 *Confusion Matrix*

Confusion Matrix adalah sebuah cara mengevaluasi metode klasifikasi akurasi dari hasil klasifikasi. Akurasi dari klasifikasi akan berpengaruh terhadap performa dari suatu klasifikasi. Analisa dilakukan dengan membandingkan sebuah matrik dari prediksi kelas asli dari data [17].

Confusion Matrix akan menghitung akurasi pada sistem *data mining* atau sistem pendukung keputusan (DSS) melakukan analisis terhadap *classifier* apakah *classifier* tersebut baik atau tidak dalam mengenali kelas yang berbeda [21].

2.12 Penelitian Terdahulu

Tabel 2. 1 Penelitian Terdahulu

No.	Judul Jurnal	Nama Jurnal	Penulis	Permasalahan	*Metode*	Kesimpulan	Adopsi dari Penelitian
1	*Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques.*	*DATA ANALYTICS 2017 : The Sixth International Conference on Data Analytics*	Elshrif Elmurngi, Abdelouahed Gherbi	Bagaimana cara mengekstrak emosi di dalam opini, dan bagaimana mendeteksi ulasan positif palsu dan ulasan negatif palsu dari ulasan opini.	*NB, K-NN-IBK, K*, SVM, DT-J48.*	Akurasi tertinggi diperoleh dengan algoritma SVM	Pemahaman pendeteksian ulasan palsu

No.	Judul Jurnal	Nama Jurnal	Penulis	Permasalahan	*Metode*	Kesimpulan	Adopsi dari Penelitian
	Vol. 5 No. 1 Tahun 2017	*Detecting*					
2	Deteksi Fake Reviews Menggunakan Support Vector Machine. Vol. 6 No. 2 Tahun 2019	*E-Proceeding of Engineering.*	Bety Elysabeth Pasaribu, Anisa Herdiani, Widi Astuti	Bagaimana cara mendeteksi fake review pada kumpulan review produk	*SVM*	Mendapatkan akurasi sebesar 74.46%	Pemahaman pendeteksian ulasan palsu dan pemahan algoritma *Support Vector Machine*
3	*Fake review detection from a product review using modified method of iterative computation framework.* Vol. 58 No.1 Tahun 2016	*MATEC Web of Conferences*	Eka Dyar Wahyuni and Arif Djunaidy	Mendeteksi ulasan palsu untuk suatu produk dengan menggunakan properti teks dan rating dari sebuah ulasan.	*Iterative Vomputing Framework (ICF). Dan ICF++*	Akurasi ICF ++ lebih tinggi dibandingkan ICF	Pemahaman pendeteksian ulasan palsu
4	*Fake Reviews of Customer Detection Using Machine Learning Models.* Vol. 29 No. 6	*International Journal of Advanced Science and Technology*	D. Vijia, Nikhil Asawa, Tanay burrej a,	Kepalsuan ulasan tentang produk dan layanannya	*Logistic regression,* *MultinomialNB,* *Decision Tree,*	Akurasi tertinggi didapatkan dengan menggunakan algoritma SVM sebesar 84.88%	Pemahaman pendeteksian ulasan palsu, algoritma *Logistic Regression,* dan algoritma *SVM*

No.	Judul Jurnal	Nama Jurnal	Penulis	Permasalahan	*Metode*	Kesimpulan	Adopsi dari Penelitian
	Tahun 2020				*XG Boost Classifier*, *Ada Boost Classifier*, *SVM*		
5	Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan *Naive Bayes* dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile). Vol. 3, No. 2 Tahun 2017	*Systemic: Information System and Informatics Journal*	Ferly Gunawan, M. Ali Fauzi, Putra Pandu Adikara	Memisahkan ulasan positif dan negatif aplikasi BCA Mobile	*Multinomial Naive Bayes-Levenshtein Distance*	Metode klasifikasi Naive Bayes dan Levenshtein distance dapat diterapkan pada analisis sentimen ulasan aplikasi mobile dengan nilai akurasi 94.4%	Pemahaman algoritma *Naïve Bayes*
6	*A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia.* Vol. 8 No. 5 Tahun 2019	*International Journal of Advanced Trends in Computer Science and Engineering*	Rein Rachman Putra, Monika Evelin Johan, Emil Robert Kaburuan	Perusahaan Fintech perlu mengetahui pendapat pengguna mereka secara Realtime untuk menghadapi pesaing.	*Naïve Bayes*	Akurasi dataset A sebesar 78% dan dataset B 74%. Proses *cleansing* pada review penggguna tidak berkontribusi secara signifikan dalam	Pemahaman algoritma *Naïve Bayes*

No.	Judul Jurnal	Nama Jurnal	Penulis	Permasalahan	*Metode*	Kesimpulan	Adopsi dari Penelitian
						penelitian ini.	
7	Perbandingan Analisis Diskriminan Dan Regresi Logistik Untuk Mengklasifikasikan Kelayakan Visitasi Pelamar Bidikmisi. Vol. 9 No. 1 Tahun 2020	E-Jurnal Matematika	Nisa Hidayati, I Komang Gde Sukarsa, Desak Putu Eka Nilakusmawati	Membandingkan analisis diskriminan dan regresi logistik untuk mengklasifikasikan kelayakan kunjungan pelamar Bidikmisi berdasarkan akurasi klasifikasi	*Logistic Regression*	Hasil dari Regresi Logistik 98.21% lebih besar dibandingkan analisis diskriminan 96.64%	Pemahaman Algoritma Regresi Logistik

Pada analisis sentimen ini, terdapat tujuh (7) jurnal yang digunakan sebagai acuan utama dalam menjalankan penelitian ini yaitu:

1. Jurnal *Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques* digunakan sebagai pemahaman pendeteksian ulasan palsu. Salah satu masalah utama yang dihadapi analisis sentimen adalah bagaimana mengekstrak emosi di dalam opini, dan bagaimana mendeteksi ulasan positif palsu dan ulasan negatif palsu dari ulasan opini [22].
2. Jurnal Deteksi Fake Reviews Menggunakan *Support Vector Machine* sebagai pemahaman pendeteksian ulasan palsu dan pemahan algoritma *Support Vector Machine.* Calon pembeli yang semakin selektif dalam membeli barang di *e-commerce* bergantung kepada ulasan yang diberikan oleh pembeli lainnnya untuk menentukan keputusan membeli suatu

produk. Penelitian deteksi *fake review* pada produk menggunakan algoritma *Support Vector Machine* [23].

3. Jurnal *Fake review detection from a product review using modified method of iterative computation framework* sebagai Pemahaman pendeteksian ulasan palsu. Ulasan dapat dijadikan sumber informasi. Misalnya, perusahaan dapat menggunakan ulasan sebagai keputusan desain produk atau layanan, sedangkan pelanggan dapat menggunakan ulasan untuk memutuskan membeli atau menggunakan produknya. Sayangnya ulasan banyak disalahgunakan oleh pihak-pihak tertentu dengan membuat ulasan palsu dengan tujuan menaikkan popularitas atau menjatuhkan produk dan toko [6].
4. Jurnal *Fake Reviews of Customer Detection Using Machine Learning Models* sebagai pemahaman pendeteksian ulasan palsu, algoritma *Logistic Regression* dan algoritma *Support Vector Machine*. Penelitian ini mendeteksi ulasan palsu dengan *Machine Learning* kemudian dilakukan klasifikasi terhadap data dengan membandingkan model klasifikasi *Logistic Regression, MultinomialNB, Decision Tree, XG Boost Classifier, Ada Boost Classifier,* dan *SVM. SVM* menghasilkan akurasi paling besar yaitu 84.88% [5].
5. Jurnal Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile) sebagai pemahaman algoritma *Naïve Bayes.* Ulasan ditujukan untuk mengevaluasi dan meningkatkan kualitas produk

kedepannya agar lebih baik. Klasifikasi sentimen dibagi ke kelas positif dan negatif. Penelitian ini menggunakan algoritma *Naïve Bayes* dengan hasil nilai akurasi sebesar 94.4% [24].

6. Jurnal *A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in* Indonesia sebagai Pemahaman algoritma *Naïve Bayes.* Perusahaan *Fintech* perlu mengetahui pendapat pengguna secara *Realtime* untuk menghadapi pesaing di pasar. Analisis sentimen dilakukan pada penelitian ini menggunakan algoritma *Naïve Bayes* untuk mengklasifikasikan ulasan pengguna berdasarkan subjek. Akurasi yang dihasilkan adalah 78% untuk data bahasa Inggris dan 75% untuk data bahasa Indonesia [25].
7. Jurnal Perbandingan Analisis Diskriminan Dan Regresi Logistik Untuk Mengklasifikasikan Kelayakan Visitasi Pelamar Bidikmisi sebagai pemahaman Algoritma Regresi Logistik. Penelitian ini membandingkan analisis diskriminan dan regresi logistik untuk mengklasifikasikan kelayakan kunjungan pelamar Bidikmisi berdasarkan akurasi klasifikasi. Algoritma Regresi Logistik menghasilkan akurasi sebesar 98.21% dibandingkan diskriminan sebesar 94.64%.

UNIVERSITAS MULTIMEDIA NUSANTARA