BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Untuk melaksanakan penelitian ini acuan dari beberapa penelitian terdahulu menjadi sangat penting dalam melakukan sebuah penelitian dengan tujuan untuk mengetahui hubungan antara penelitian yang akan dilakukan dengan penelitian terdahulu, sehingga dengan menambahkan acuan tersebut dapat menghindari adanya suatu duplikasi dalam penelitian yang akan dilakukan. Analisis sentimen merupakan bidang penelitian berkelanjutan yang berada diantara berbagai bidang seperti *Data Mining, Natural Language Processing* dan *Machine Learning* yang berfokus pada ekstraksi sentimen dalam sebuah kalimat berdasarkan isinya.

Studi komputasional dari opini pengguna sosial media, sentimen dan emosi melalui entitas dan atribut yang dimiliki dan di ekspresikan dalam bentuk teks dewasa ini banyak dilakukan karena topik ini sangat menarik untuk dibahas. Pengembangan penelitian dan pengklasifikasikan sentimen analisis menjadi alternatif dan bahan evaluasi yang baik bagi pihak terkait. Hal ini terbukti dengan banyaknya riset atau penelitian di bidang tersebut. Penelitian terdahulu mengenai analisis sentimen telah dilakukan oleh peneliti pada berbagai media sosial dan objek yang berbeda.

Penelitian yang berhasil peneliti temukan adalah penelitian analisis sentimen yang dilakukan oleh Servasius Dwi Harjiatno (2019). Pada penelitian ini, sistem akan melakukan klasifikasi terhadap opini masyarakat di media sosial *Twitter* terhadap tokoh publik menjelang pemilu 2019, yaitu Jokowi dan Prabowo yang diklasifikasikan dalam 5 kelompok sentiment yaitu cinta, marah, senang, sedih, dan takut. Metode yang digunakan adalah *Multinominal Naïve Bayes.*

Yang kedua, penelitian yang dilakukan oleh Awwaluddin (2021). Pada penelitian ini menganalisis sentiment negatif, positif, dan netral dari *Tweet*

nasabah pada akun *Twitter* BNICustemoreCare pada pelayanan Internet Banking menggunakan metode *Naïve Bayes*. Dan *Tweet* yang digunakan untuk analisis adalah data tahun 2020.

Yang ketiga penelitian yang dilakukan oleh Tiara Ramadhani (2021). Penelitian ini menganalisis tentang opini masyarakat terhadap acara TV yang dianalisis sentiment menjadi 3 (Tiga) sentiment yaitu negatif, positif, dan netral. Metode yang digunakan adalah metode *K-Nearest Neighbour.*

Yang ketiga penelitian yang dilakukan oleh Pikir Claudia Septiani Gulo (2021). Analisis Sentimen tentang pengaruh kuliah online selama pandemi covid-19. Analisis yang dilakukan hanya menganalisis data *Tweet* berbahasa Indonesia dengan mengelompokan sentiment menjadi positif, negatif, dan netral. Metode yang digunakan adalah metode *Naïve Bayes.*

Yang keempat penelitian yang dilakukan oleh Farahdiva Assyfa Andrin (2021). Yaitu menganalisa anemo masyarakat terhadap isu kecurangan paska pilpers 2019. Dari penelitian ini menghasilkan data training dan testing yang akan dilakukan klasifikasi menggunakan pohon keputusan dengan Algoritma C4.5 untuk menentukan Sentiment Analysis. Hasil pengujian di peroleh dengan akurasi yang cukup tinggi yaitu 80%. Analisa sentiment anemo masyarakat terhadap isu kecurangan paska pilpres 2019 tersebut fokusnya ada pada tingkat polaritas respon atau pendapat kedalam kategori positif dan negatif.

Pada penelitian ini, penulis akan menganalisis sentiment positif, negatif, dan netral dari t*weet* berbahasa Indonesia dengan kata kunci Telkomsel. Data yang digunakan adalah *Tweet* bulan Januari 2022 sampai April 2025 atau tidak lebih dari 2000 data. Data diambil secara acak baik dari user biasa ataupun media online di *X* dan analisis sentimen menggunakan metode *Naïve Bayes.*

Tabel 2.1 Tinjauan Pustaka

No.	Penulis	Obyek Penelitian	Metode	Keterangan
1	Servasius Dwi Harijiatno (2019)	Analisis Sentimen Pada *Twitter* Menggunakan *Multinominal Naïve Bayes*	*Naïve Bayes*	Hasil uji akurasi tanpa menggunakan *k-Fold Cross Validation* menghasilkan akurasi yang lebih besar, yaitu 72.491% dibandingkan dengan semua *k-Fold Validation* dimana 3 *Fold* menghasilkan akurasi 71.601%, *5-Fold* menghasilkan akurasi 70.72%, dan *10-Fold* menghasilkan akurasi 70.68%
2	Awwaluddin (2021)	Analisis Sentimen Pada Akun *Twitter* BNI *Customer Care* Menggunakan Metode *Naïve Bayes Classifier (NBC)*	*Naïve Bayes*	Berhasil membuat aplikasi yang mampu mengolah data mentah menjadi sebuah informasi yang berguna untuk meningkatkan pelayanan BNI
3	Tiara Ramadhani (2021)	Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Mayarakat Pada Media Sosial *Twitter* Menggunakan Metode *K-Nearst Neighbour*	*K-Nearst Neighbour*	Hasil akurasi pengujian klasifikasi pada metode KKN menghasilkan akurasi sebesar 72.56% pada K=3

Tabel 2.2 Lanjutan Tinjauan Pustaka

No.	Penulis	Obyek Penelitian	Metode	Keterangan
4	Pikir Claudia Septiani Gulo (2021)	Analisis Sentimen Kuliah Online Selama Pandemi Covid-19 Menggunakan Algoritma *Naïve Bayes*	*Naïve Bayes*	Hasil analisis nilai *Precision* sebesar 79%, nilai *Recall* sebesar 80% dan nilai F1-*Score* sebesar 79%
5	Dita Widianti (Diusulkan)	Analisis Sentimen Pada *Twitter* Mengenai Pelayanan *Provider* Telkomsel Menggunakan Metode *Naïve Bayes*	*Naïve Bayes*	Akan menghasilkan akurasi yang dapat diperoleh dalam analisis sentiment pada media sosial X tentang layanan *provider* telkomsel dengan mentode *Naïve Bayes*

2.2 Dasar Teori

2.2.1 Analisis Sentimen

Analisis sentimen adalah bidang studi yang menganalisis pendapat orang, opini, evaluasi, penilaian, sikap dan emosi terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik dan atribut (Liu, 2012). Sentimen menurut Kamus Besar Bahasa Indonesia (KBBI) adalah:

1. Pendapat atau pandangan yang didasarkan pada perasaan yang berlebih-lebihan terhadap sesuatu (bertentangan dengan pertimbangan pikiran). Contoh: keputusan yang dihasilkan akan tidak adil jika disertai rasa sentimen pribadi.

2. Emosi yang berlebihan. Contoh: rasa sentimen sebagai bangsa Indonesia akan tumbuh kuat jika kita jauh dari negeri ini.

3. Iri hati; tidak senang; dendam.

4. Reaksi yang tidak menguntungkan. Contoh: penurunan harga saham hanya disebabkan oleh sentimen pasar

Analisis sentimen dilakukan untuk melihat pendapat terhadap sebuah masalah, atau dapat juga digunakan untuk identifikasi kecenderungan hal yang sedang menjadi topik pembicaraan. Analisis sentimen pada penelitian ini adalah proses pengelompokan *Tweet* kedalam tiga sentiment yaitu positif, negatif, dan netral.

2.2.2 X

X atau yang mulanya disebut *Twitter* adalah layanan jejaring sosial dan mikroblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter, akan tetapi pada tanggal 07 November 2017 bertambah hingga 280 karakter yang dikenal dengan sebutan kicauan *(Tweet*). X dibentuk pada tahun 2006 oleh Jack Dorsey. X berbasis di San Brunomor, California dekat San Francisco, dimana situs ini pertama kali dibuat

(Rezeki, Restiviani, & Zahara, 2020). *Twitter* berganti nama menjadi X pada Juli 2023. X telah menjadi salah satu dari sepuluh situs yang paling sering dikunjungi di Internet, dan dijuluki dengan pesan singkat dari Internet. X mengalami pertumbuhan yang pesat dan dengan cepat meraih popularitas di seluruh dunia. Hingga Mei 2015, X telah memiliki lebih dari 500 juta pengguna, 302 juta di antaranya adalah pengguna aktif. Sedangkan di Indonesia, Menurut laporan *We Are Social*, jumlah pengguna X mencapai 18,45 juta pada 2022. Angka tersebut menempatkan Indonesia di peringkat ke lima negara pengguna X terbesar di dunia. (Rizaty, 2022).

2.2.3 *Text Mining*

Text mining merupakan suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan *tools analysis* yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi (Feldman & Sanger, 2007). *Text mining* dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokkan dan menganalisa *unstructured text* dalam jumlah besar. Dalam memberikan solusi, *text mining* mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *Data mining, Information Retrieval, Statistik dan Matematik, Machine Learning, Linguistic, Natural Languange Processing (NLP),* dan *Visualization.*

Tujuan dari *Text Mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen, tetapi tujuan utama *text mining* adalah mendukung proses *knowledge discovery* pada koleksi dokumen yang besar. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks (*text categorization*) dan pengelompokkan teks (*text clustering*) (Feldman & Sanger, 2007).

2.2.4 *Scrapping*

Scraping merupakan teknik atau metode otomatisasi untuk mengekstrak data dari sebuah *website*, *database*, aplikasi *enterprise*, atau sistem l*egacy* yang kemudian dapat menyimpannya ke dalam sebuah file dengan format tabular atau *spreadsheet*. Kebanyakan data pada *website* merupakan data tidak terstruktur dalam format HTML yang kemudian diubah menjadi data dengan format terstruktur ke dalam *spreadsheet* atau *database* sehingga dapat dimanipulasi. Manfaat menggunakan data *scraping* adalah efisiensi waktu dan tenaga.

2.2.5 *Laxicon Based*

Metode *Lexicon Based* adalah salah satu cara umum dalam menganalisis sentimen pada media sosial. Metode ini menggunakan kamus sebagai sumber bahasa yang berfungsi untuk mengklasifikasikan sentimen dari setiap opini sehingga kalimat sentimen dapat diklasifikasikan ke dalam kelas negatif, positif, dan netral.

Tahap pembobotan *Lexicon Based* dalam analisis sentimen bertujuan untuk memberikan nilai numerik pada kata-kata berdasarkan sentimen yang terkandung di dalamnya. Nilai ini kemudian digunakan untuk menghitung sentimen keseluruhan dari sebuah kalimat. Kata yang teridentifikasi dalam kamus *lexicon* akan dihitung skornya sesuai dengan jumlah kata pada setiap teks atau kalimat.

$$Spositive \sum_{i\epsilon t}^{n} Positive\ Score\ i \qquad (1)$$

$$Sne\text{gative} \sum_{i\epsilon t}^{n} Negative\ Score\ \text{i} \qquad (2)$$

Dimana ($Spositive$) adalah bobot dari kalimat yang didapatkan melalui penjumlahan n skor polaritas kata opini positif dan ($Snegative$) adalah bobot dari kalimat yang didapatkan melalui penjumlahan n skor polaritas kata opini negatif. Dari persamaan nilai sentimen dalam satu kalimat maka diperoleh persamaan 3 untuk menentukan orientasi sentimen dengan perbandingan jumlah nilai positif,

negatif dan netral.

$$Sentence_{sentiment} \begin{Bmatrix} positive\ if\ S_{positive} > S_{negative} \\ neutral\ if\ S_{positive} = S_{negative} \\ negative\ if\ S_{positive} < S_{negative} \end{Bmatrix} \quad (3)$$

Jika dalam suatu teks memiliki jumlah kata positif lebih banyak dari kata negatif, maka data teks tersebut akan dilabeli sentimen positif. Jika dalam suatu teks memiliki jumlah kata positif lebih sedikit dari kata negatif, maka data teks tersebut akan dilabeli sentimen negatif. Jika dalam suatu teks memiliki jumlah kata positif sama dengan kata negatif, maka data teks tersebut akan dilabeli sentimen netral.

2.2.6 *Naive Bayes Classifier*

Metode *Naïve Bayes* merupakan salah satu metode pembelajaran mesin 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* bekerja sangat baik dibanding dengan model *classifier* lainnya. Hal ini dibuktikan oleh Xhemali, Hinde Stone dalam jurnalnya "*Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages*" bahwa "*Naïve Bayes Classifier* memiliki tingkat akurasi yang lebih baik dibanding model *classifier* lainnya". Keuntungan penggunan metode ini hanya membutuhkan jumlah data pelatihan (*training data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variable independent, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

Adapun jenis-jenis algoritma *Naïve Bayes* adalah :

1. *Multinominal Naïve Bayes*

Algoritma *Multinominal Naïve Bayes* adalah algoritma yang paling sering digunakan untuk klasifikasi teks. Klasifikasi ini cocok untuk fitur dalam bentuk frekuensi atau hitungan kata.

2. *Bernoulli Naïve Bayes*

Algoritma *Bernoulli Naïve Bayes* adalah algoritma yang digunakan untuk fitur biner. Algoritma ini cocok untuk dokumen pendek atau klasifikasi spam.

3. *Gaussian Naïve Bayes*

Algoritma *Gaussian Naïve Bayes* digunakan jika fitur memiliki distribusi kontinu yang menyerupai distribusi normal.

Pada penelitian ini, jenis metode naive bayes yang digunakan adalah Multinomial Naive Bayes. Metode Multinominal Naive Bayes adalah salah satu variasi algoritma Naive Bayes. Algoritma klasifikasi berdasarkan Teorema Bayes ini ideal untuk data diskrit dan biasanya digunakan dalam permasalahan klasifikasi teks. Algoritma ini memodelkan frekuensi kata sebagai hitungan dan mengasumsikan setiap fitur atau kata terdistribusi secara multinomial. MNB banyak digunakan untuk tugas-tugas seperti mengklasifikasikan dokumen berdasarkan frekuensi kata, seperti dalam deteksi email spam.

Dalam Multinomial Naive Bayes, kata "Naif" berarti metode ini mengasumsikan semua fitur, seperti kata-kata dalam kalimat, bersifat independen satu sama lain. "Multinomial" mengacu pada seberapa sering sebuah kata muncul atau seberapa sering sebuah kategori muncul. Metode ini bekerja dengan menggunakan jumlah kata untuk mengklasifikasikan teks.

Algoritma Multinominal Naive Bayes pada program ini menggunakan library Skicit-Learn atau disebut juga dengan Sklearn.

Secara umum, teorema Bayes dapat dinyatakan secara matematis dalam persamaan [4].

$$P\ (H|\mathrm{X} = \frac{P(H)P(X|\mathrm{H})}{P(X)} \quad (4)$$

X= data dengan kelas tidak dikenal

H= hipotesis data X adalah kelas khusus

P(H(X)= probabilitas hipotesis H didasarkan pada kondisi X

P(X) = Probabilitas X

2.2.7 *Text Preprocessing*

Text preprocessing proses mengubah data mentah menjadi data yang sesuai dengan prosedur mini yang akan dilakukan dan merupakan tahap yang paling penting dalam *data mining*. Dalam tahap ini dilakukan proses pembersihan pada data yang masih kotor (Fikri, M, & Azhar, 2020). Tahapan yang dilakukan di *text processing* adalah :

1. *Cleaning*

Tahapan *cleaning* yaitu penghapusan karakter yang tidak penting seperti angka, *hastag, symbol*, http, dan karakter lainnya.

2. *Case folding*

Pada tahap *case folding* dilakukan perubahan setiap dokumen yang memiliki huruf kapital menjadi huruf kecil.

3. Normalisasi Kata

Normalisasi kata adalah proses dari memperbaiki dari kata-kata yang mengalami salah pengejaan maupun disingkat dalam bentuk tertentu.

4. *Tokenizing* (Tokenisasi)

Tokenisasi merupakan tahapan melakukan pemisahan terhadap kalimat untuk menjadi per kata. Tahap ini adalah proses pemecahan kata pada kalimat atau pada suatu dokumen. Dibawah ini adalah contoh tokenization dalam bentuk tabel.

5. *Stop Removal/Filtering*

Stopword Removal merupakan tahapan untuk mendapatkan kata – kata yang penting untuk menjadi indeks pada setiap dokumennya. Lanjutan dari tahapan *tokenizing* adalah tahapan *filtering* yang digunakan untuk mengambil kata-kata yang penting dari hasil token tadi. Kata umum yang biasanya muncul dan tidak memiliki makna disebut dengan *stopword*. Misalnya penggunaan kata penghubung seperti dan, yang, serta, setelah, dan lainnya. Penghilangan *stopword* ini dapat mengurangi ukuran index dan waktu pemrosesan.

6. *Stemming*

Tahap ini diperlukan untuk memperkecil jumlah indeks yang berbeda dari satu data sehingga sebuah kata yang memiliki *suffix* maupun *prefix* akan kembali ke bentuk dasarnya. Selain itu juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk yang berbeda karena mendapatkan imbuhan yang berbeda pula.

2.2.8 *Google Collaboratory*

Google Colabaratory atau sering disingkat Google Colab adalah platfrom *cloud* yang disediakan oleh Google untuk menulis dan menjalankan kode *Python* melalui browser tanpa memerlukan konfigurasi tambahan yang memanfaatkan infrastruktur *cloud* Google dan memberikan lingkungan pengembangan yang kuat dengan akses ke GPU & TPU (Unit Pemrosesan Tensor) secara gratis (Nazar, 2024).

2.2.9 *Python*

Python merupakan bahasa pemrograman tingkat tinggi yang mendukung pemrograman berorientasi objek. *Python* sering digunakan dalam aplikasi web, pengembangan perangkat lunak, ilmu data, dan mechine learning. *Python* merupakan sebuah bahasa pemrograman yang bersifat *general-purpose* untuk mendukung pemrograman berorientasi objek. *Python* memeiliki perbedaan dengan bahasa pemrograman yang lain yaitu dalam penulisan sintaks yang mudah dipahami dan bersifat ekspresif.

2.2.10 *Mechine Learning*

Machine learning merupakan kecerdasan buatan (*Articial Intelligence*) yang bisa membuat sistem mempunyai keahlian belajar sendiri secara otomatis serta tingkatkan kemampuannya berdasarkan pengalaman tanpa perlu diprogram oleh manusia (Mayang, 2021). *Machine Learning* juga merupakan salah satu cabang dari ilmu Kecerdasan Buatan, khususnya yang mempelajari tentang bagaimana komputer mampu belajar dari data untuk meningkatkan kecerdasannya, *Machine learning* memiliki fokus pada pengembangan sebuah sistem yang mampu belajar sendiri untuk memutuskan sesuatu, tanpa harus berulang kali diprogram oleh manusia. Dengan metode tersebut, mesin tidak hanya bisa menemukan aturan untuk perilaku optimal dalam pengambilan keputusan, namun juga bisa beradaptasi dengan perubahan yang terjadi (Wahyono, 2018).