BAB 2
LANDASAN TEORI

2.1 Penelitian Terdahulu

Penelitian terdahulu juga pernah dilakukan oleh Azwan Triyadi bersama rekan-rekannya dengan menerapkan SMOTE kedalam model klasifikasi nya, pada penelitannya menggunakan 3 Model untuk klasifikasi dari Naïve Bayes yaitu Complement Naïve Bayes, Multinomial Naïve Bayes, dan Gaussian Naïve Bayes. Dataset yang digunakan dalam penelitiannya mencakup tweet dengan tagar bahasa Inggris "neuralink" di Twitter dan diambil antara 1 Januari 2023 hingga 27 Juni 2023, menunjukan Complement Naïve Bayes mencapai Accuracy 81%, Precision 82%, Recall 81%, dan F1-score 79%. Sementara Multinomial Naïve Bayes mencapai Accuracy 80%, Precision 83%, Recall 80%, dan F1-score 79%, dan Gaussian Naïve Bayes yang mencapai nilai Accuracy 75%, Precision 78%, Recall 75%, dan F1-score 72%[19].

Selain itu penelitian serupa juga dilakukan oleh Muhammad Yusran dan rekan-rekan nya untuk membandingkan antara 2 metode klasifikasi Multinomial Naïve Bayes dan Bernoulli Naïve Bayes. Dataset nya diambil dari Twitter berupa twit berbahasa Indonesia dengan tagar "Kurikulum Merdeka" dari 21 Juli 2022 hingga 17 November 2022. Hasiluji dari penelitian nya Multinomial Naïve Bayes mendapat poin Accuracy 98,88%, Recall 98,13%, Precision 99,05%, dan F1-score 98,51%, sementara Bernoulli Naïve Bayes mendapat Accuracy 94,81%, Recall 87,85%, Precision 98,94%, dan F1-score 93,06%. Bisa disimpulkan dari perbandingan kedua metode klasifikasi diatas Multinomial Naïve Bayes lebih baik daripada Bernoulli Naïve Bayes[18].

Penelitian selanjutnya juga pernah dilakukan oleh Merve bersama rekan-rekannya, pada penelitiannya menggunakan banyak model klasifikasi yang membandingkan antar klasifikasi 4 Naïve Bayes dengan model klasifikasi seperti Random Forest, KNN, Linear SVC, Multi Layer Perceptron (MLP), Linear Regression, dan Decision Tree. Dataset yang dikumpulkan diambil dari BBC News Corpus dan dataset ini terdiri dari lima kategori berbeda : teknologi, bisnis, olahraga, hiburan, dan politik dari total ke lima kategori mendapatkan sebanyak 2225 data, sehingga menunjukkan beberapa nilai yang di dapatkan dari model klasifikasi nya seperti Complement Naïve Bayes Accuracy 98,31%, Precision

98,31%, Recall 98,31% dan F1-score 98,31%. Multinomial Naïve Bayes Accuracy 98,20%, Precision 98,20%, Recall 98,20% dan F1-score 98,20%. Bernoulli Naïve Bayes Accuracy 96,67%, Precision 97,01%, Recall 96,74% dan F1-score 96,77%. Gaussian Naïve Bayes Accuracy 92,92%, Precision 93,08%, Recall 92,92% dan F1-score 92,93%, setelahnya mereka juga mendapatkan nilai dari model klasifikasi dari Naïve Bayes seperti MLP Classifier Accuracy 98,31%, Precision 98,32%, Recall 98,31% dan F1-score 98,31%. Linear SVC Accuracy 97,97%, Precision 97,98%, Recall 97,97% dan F1-score 97,98%. Random Forest Accuracy 97,86%, Precision 97,91%, Recall 97,86% dan F1-score 97,87%. Logistic Regression Accuracy 97,52%, Precision 97,52%, Recall 97,52% dan F1-score 97,52%. Decision Tree Accuracy 82,69%, Precision 82,84%, Recall 82,69% dan F1-score 82,72%. Maka dapat disimpulkan bahwa model klasifikasi Naïve Bayes terbukti tangguh dalam teks klasifikasi berita[20].

Tabel 2.1 menjelaskan kerangka penelitian sejenis yang digunakan dalam mencari referensi pendukung dalam membuat penelitian ini. Berikut dibawah ini lampiran tabel nya:

Tabel 2.1. Kerangka Penelitian dari Tiga Studi Terdahulu

Komponen	**Penelitian 1: Azwan Triyadi dkk. (Neuralink)**	**Penelitian 2: Muhammad Yusran dkk. (Kurikulum Merdeka)**	**Penelitian 3: Merve dkk. (BBC News Corpus)**
Sumber Data	Tweet Inggris bertagar "neuralink" (1 Jan–27 Jun 2023).	Tweet Indonesia bertagar "Kurikulum Merdeka" (21 Jul–17 Nov 2022).	BBC News Corpus (2225 artikel, 5 kategori: teknologi, bisnis, olahraga, hiburan, politik).
Metode Labeling	Label otomatis menggunakan model RoBERTa.	Label manual (metode labeling tidak dijelaskan rinci pada sumber).	Dataset sudah memiliki label kategori (news categories).
Pra-pemrosesan	Pembersihan teks + ekstraksi fitur menggunakan TF-IDF.	Pembersihan teks + seleksi fitur menggunakan Query Expansion Ranking (QER).	Pembersihan teks + ekstraksi fitur (TF-IDF / Vector Space Model).

Penanganan Imbalance	SMOTE digunakan untuk menyeimbangkan kelas.	Tidak disebutkan SMOTE; fokus pada seleksi fitur QER.	Tidak disebutkan imbalance; dataset relatif seimbang.
Pembagian Data	Train–test split 80% : 20%.	Pembagian data latih–uji (rasio tidak dijelaskan).	Pembagian data latih–uji standar (umumnya 80:20).
Model Klasifikasi	Complement NB, Multinomial NB, Gaussian NB.	Multinomial NB dan Bernoulli NB.	CNB, MNB, BNB, GNB, Random Forest, KNN, Linear SVC, MLP, Logistic Regression, Decision Tree.
Evaluasi Kinerja	Accuracy, Precision, Recall, F1-score.	Accuracy, Precision, Recall, F1-score.	Accuracy, Precision, Recall, F1-score.
Hasil Utama	CNB terbaik (Accuracy 81%).	MNB unggul (Accuracy 98.88%).	NB dan MLP menunjukkan performa tertinggi (Accuracy $\approx$ 98%).
Framework Metodologis	TF-IDF, SMOTE, Naïve Bayes variants, RoBERTa labeling, evaluasi metrik standar.	Preprocessing, QER feature selection, MNB vs BNB, evaluasi metrik.	TF-IDF, multi-model comparison, multi-class classification, evaluasi metrik.

2.2 Lowongan Lapangan Pekerjaan

Lowongan kerja seharusnya memberikan informasi yang jelas agar pencari kerja mudah memahami jenis pekerjaan yang ditawarkan. Hal ini sangat penting bagi mereka yang sedang mencari pekerjaan atau belum memiliki perencanaan karir tetapi kenyataannya banyak lowongan yang tidak mencantumkan informasi lengkap sehingga membuat pencari kerja kebingungan, meskipun informasi lowongan kerja kini mudah diakses secara online namun karena banyaknya pilihan pekerjaan justru dapat membuat mereka semakin sulit menentukan pilihan. Selain itu tidak semua lowongan menampilkan keterangan penting seperti syarat pekerjaan atau informasi Perusahaan yang mengakibatkan pencari kerja kesulitan dalam memahami kondisi lapangan kerja saat ini oleh karena itu, diperlukan penyajian informasi lowongan yang lebih lengkap dan mudah dipahami[21].

2.3 Google Play Store

Google Play Store adalah aplikasi Android resmi yang dikembangkan oleh Google Inc. dan memungkinkan pengguna mendapatkan berbagai layanan digital, seperti aplikasi, game, film, dan e-book. Aplikasi ini terus berkembang dan berinovasi dalam berbagai macam-macam aplikasi[22].Fitur yang ada didalam aplikasi Google Play Store ini adalah dapat memberikan rating dan ulasan dari user pengguna nya dan dapat memberikan opini dari aplikasi yang telah mereka gunakan[23].Google Play Store merupakan app store utama yang berjalan pada sistem operasi Android yang penggunanya dapat mencari dan mengunduh aplikasi yang dikembangkan menggunakan Android SDK (Software Development Kit) oleh Google[24].

2.4 Google Play Store Review

Google Play Store menyediakan fitur ulasan sebagai ruang bagi pengguna untuk menyampaikan keluhan, saran, serta penilaian terhadap aplikasi yang telah mereka gunakan. Setiap halaman aplikasi memiliki kolom komentar yang berfungsi sebagai wadah bagi pengguna untuk memberikan masukan saran maupun kritik. Informasi dari ulasan tersebut menjadi sumber berharga bagi pengembang dalam mengevaluasi kinerja aplikasi dan menetapkan langkah perbaikan yang diperlukan berdasarkan rekomendasi pengguna. Melalui ulasan pengguna, kelebihan dan kekurangan suatu aplikasi dapat dianalisis secara lebih mendalam untuk membantu pengembang meningkatkan kualitas layanan dan mempermudah pengalaman pengguna. Maka, Google Play Store berperan penting sebagai platform yang menyediakan penilaian jujur dari pengguna terhadap kualitas setiap aplikasi yang tersedia[25].

2.5 Glints : TapLoker

Glints adalah platform aplikasi mobile yang berfokus pada jasa penyedia lowongan kerja dan pengembangan karir. Aplikasi ini menyediakan informasi tentang lowongan kerja, termasuk pekerjaan full-time, internship, dan freelance, serta kelas untuk belajar tentang memulai karir dan pengembangan diri[26].

Gambar 2.1. Logo Glints: TapLoker
(Sumber: play.google.com, 2025)

Gambar 2.1 diatas merupakan visualisasi tampilan logo aplikasi Glints: TapLoker yang dapat diunduh pada Google Playstore. Glints dapat membantu pencari kerja menemukan informasi yang mereka butuhkan. Aplikasi ini memungkinkan pengguna untuk melamar pekerjaan dengan didalamnya terdapat fitur seperti ”Lamar hanya dengan 1x Tap”, para pengguna memungkinkan dapat berbicara langsung dengan HRD, Filter Lowongan yang Luas memungkinkan mencari pekerjaan berdasarkan gaji, tempat, keahlian, jenis kontrak (full-time, part-time, atau freelance), dan lokasi kerja terdekat. Selain itu bisa memantau status lamaran kita, pada saat HRD melihat lamaran kita maka kita akan menerima notifikasi langsung saat itu. Glints juga menjadi aplikasi dengan lowongan pekerjaan terlengkap dari berbagai Perusahaan terkenal yang ada di Indonesia dan terbukti menawakan lebih dari 50.000 posisi pekerjaan, Glints juga menawarkan tips-tips bagaimana kita terhindar dari pekerjaan palsu/mencurigakan dan menerapkan verifikasi yang sangat ketar dalam dunia bisnis[27].

Glints muncul sebagai solusi inovatif untuk tantangan pencarian kerja di kalangan generasi muda Indonesia dengan menghadirkan platform yang tidak hanya menghubungkan pencari kerja dengan lowongan, tetapi juga membantu pengembangan keterampilan dan kesiapan karier. Dengan adanya fitur konsultasi karier serta eksplorasi jalur keterampilan dan kerja sama dengan banyak perusahaan, Glints menjadi salah satu aspek penting dalam ekosistem talenta muda di Indonesia dengan potensi pertumbuhan besar karena basis penggunanya yang terus berkembang[28].

2.6 *Text Mining*

Text mining merupakan metode yang efektif untuk memproses dan menganalisis kumpulan data teks dalam jumlah besar secara otomatis. Melalui pendekatan ini, berbagai informasi penting dapat diekstraksi sehingga menghasilkan wawasan yang relevan dan dapat digunakan untuk mendukung proses pengambilan keputusan[29]. *Text mining* berperan penting dalam mengumpulkan dan mengolah dataset berskala besar dari berbagai sumber data. Teknik ini juga semakin banyak digunakan untuk menganalisis opini publik, terutama melalui pemanfaatan data yang berasal dari media sosial. Dengan kemampuannya mengekstraksi informasi secara otomatis, text mining menjadi pendekatan yang efektif untuk memahami persepsi, kecenderungan, dan pola komunikasi yang muncul dalam interaksi daring[30].

2.7 *Natural Language Processing (NLP)*

Natural Language Processing (NLP) dapat dipahami sebagai teknologi yang membuat komputer mampu membaca, memahami, dan mengolah bahasa manusia secara otomatis. Teknologi NLP sebagai bidang yang bekerja menggabungkan ilmu komputer, linguistik, dan kecerdasan buatan untuk mengubah bahasa manusia menjadi data terstruktur. NLP pada dasarnya bertujuan membantu komputer mengenali makna teks dan mengolah informasi dalam jumlah besar secara lebih efisien, karena banyak aktivitas digital diberbagai aplikasi sehari-hari, seperti ponsel pintar, asisten virtual, dan layanan pesan otomatis dan dari situ menghasilkan data teks yang sangat banyak. Teknologi NLP menjadi semakin penting untuk mengekstraksi informasi, mendukung analisis, dan mempermudah berbagai pekerjaan yang sebelumnya harus dilakukan secara manual[31].

Natural Language Processing (NLP) merupakan bidang dalam kecerdasan buatan *Artificial Intelligence* yang berfokus pada bagaimana komputer dapat memahami, mengolah, dan merespons bahasa manusia. Konsep dasar NLP berangkat dari pemahaman bahwa bahasa adalah kumpulan aturan dan simbol yang digunakan untuk menyampaikan informasi sehingga diperlukan teknologi yang mampu menerjemahkan bentuk bahasa alami ke dalam format yang dapat diproses mesin. NLP hadir sebagai solusi bagi pengguna yang tidak memiliki kemampuan atau waktu untuk mempelajari bahasa pemrograman, NLP berperan penting dalam mempermudah komunikasi antara manusia dan sistem komputer

melalui pengolahan otomatis terhadap teks maupun percakapan[32].

2.8 Analisis Sentimen

Analisis sentimen merupakan teknik mengindentifikasi dan menganalisis yang tujuannya untuk mengekstrak dan memahami opini atau perasaan yang terkandung dalam teks, proses menganalisis, memahami, dan mengekstraksi data berbentuk teks untuk mengidentifikasi sentimen atau opini yang terkandung dalam sebuah kalimat apakah bersifat positif, negatif atau netral[33]. Analisis sentimen juga menjadu solusi yang memungkinkan ekstraksi opini ringkas atau detail sentimen yang mendalam mengenai topik atau konteks apa pun dari sumber data yang sangat besar[34]. Analisis sentimen juga bertujuan untuk menganalisis pendapat, sentimen, evaluasi, penilaian, sikap dan emosi seseorang terhadap suatu entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik dan atribut mereka[35]. Analisis sentimen mencakup emosi, sikap, atau pendapat yang berkaitan dengan suatu objek atau subjek. Dalam hal ini analisis membagi berbagai jenis pendapat yang mengandung polarisasi dan dapat dengan tepat menyatakan jenis pendapat pada dokumen teks dengan jelas dibedakan dengan menggunakan model deteksi yang tepat[36].

2.9 Naïve Bayes

Naive Bayes sebagai metode klasifikasi berbasis probabilitas, Naive Bayes menerapkan prinsip-prinsip teorema Bayes dalam prosesnya[37]. Metode ini memiliki keunikan dalam hal asumsi independensi yang kuat, dimana setiap atribut klasifikasi diyakini tidak saling mempengaruhi satu sama lain. Naive Bayes bekerja dengan menganalisis cara kata-kata tertentu muncul dalam dokumen untuk menentukan kategori atau kelas dokumen tersebut. Metode Naive Bayes masih efektif, meskipun asumsi tentang independensi antara kata-kata tidak selalu terjadi, terutama ketika diterapkan pada kumpulan teks yang sangat besar[38]. Penelitian ini memanfaatkan algoritma tersebut untuk memproses dan mengklasifikasikan ulasan aplikasi Glints dari Google Play Store menjadi beberapa kategori sentimen positif, negatif, dan netral[39].

Rumus 2.1 merupakan dasar Teorema Bayes[40]:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{2.1}$$

Keterangan:

- B: Data informasi yang belum diketahui.
- A: Hipotesis mengenai data A adalah kategori spesifik.
- $P(A \mid X)$: Probabilitas dari A berdasarkan dari kondisi B.
- $P(A)$: Probabilitas dari hipotesis terkait A.
- $P(B \mid A)$: Probabilitas yang diperoleh dari B berlandaskan dari hipotesis.
- $P(B)$: Probabilitas B yang merupakan data contoh.

Rumus 2.2 merupakan Naive Bayes dalam konteks analisis sentimen:

$$P(sentimen|kata) = \frac{P(kata|sentimen) \times P(sentimen)}{P(kata)} \tag{2.2}$$

2.10 Multinomial Naïve Bayes

Algoritma ini sebagai salah satu metode dalam pemrosesan bahasa alami (NLP), Multinomial Naive Bayes menggunakan pendekatan probabilistik berdasarkan teorema Bayes. Karakteristik utama algoritma ini adalah kemampuannya menganalisis frekuensi kemunculan kata dalam dokumen. Pendekatan ini mengombinasikan dua elemen penting dalam analisisnya: identifikasi keberadaan kata dalam dokumen dan penghitungan jumlah kemunculan kata tersebut[41].

Rumus 2.3 merupakan cara perhitungan probabilitas kondisional Multinomial Naive Bayes:

$$P(A|B) = \frac{count(A,B)+1}{(\sum_{w \in V} count(A,B)) + |V|} \tag{2.3}$$

Keterangan:

- $P(A|B)$ = Probabilitas kondisional kata A pada kelas B,
- $count(A,B)$ = Jumlah kemunculan kata A dalam kategori atau kelas B,
- $\sum_{w \in V} count(A,B)$ = Total semua kata dalam kategori atau kelas B,
- $|V|$ = Semua kata unik yang ada di kategori atau kelas.

Rumus 2.4 cara untuk menghitung prior probability Multinomial Naive Bayes:

$$P(A) = \frac{N_a}{N} \tag{2.4}$$

Dimana: N_a = Jumlah dokumen dalam kelas A, N = Total jumlah dokumen

2.11 TF - IDF

Metode Term Frequency-Inverse Document Frequency (TF-IDF) menggabungkan konsep Term Frequency dan Document Frequency. Konsep Term Frequency menghitung seberapa sering suatu kata muncul dalam satu dokumen dengan mempertimbangkan frekuensi kemunculannya, dan konsep Document Frequency menghitung jumlah dokumen di mana kata tersebut muncul. Dengan metode TF-IDF, jika kata sering muncul dalam satu dokumen tetapi jarang muncul dalam dokumen umum maka kata tersebut dapat diberi nilai bobot yang lebih tinggi[17].

2.11.1 Rumus TF-IDF

Berikut dibawah ini rumus untuk menghitung TF-IDF[10].

Rumus 2.5 merupakan perhitungan dari TF (Term Frequency):

$$TF_{(t,d)} = \frac{\text{Jumlah kemunculan } t \text{ dalam } d}{\text{Total } t \text{ dalam } d} \tag{2.5}$$

Rumus 2.6 merupakan perhitungan dari IDF (Inverse Document Frequency):

$$IDF_{(t)} = \log\left(\frac{\text{Jumlah } d}{\text{Jumlah } d \text{ yang memuat } t}\right) \tag{2.6}$$

Rumus 2.7 merupakan perhitungan dari TF-IDF untuk Proses pembobotan kata:

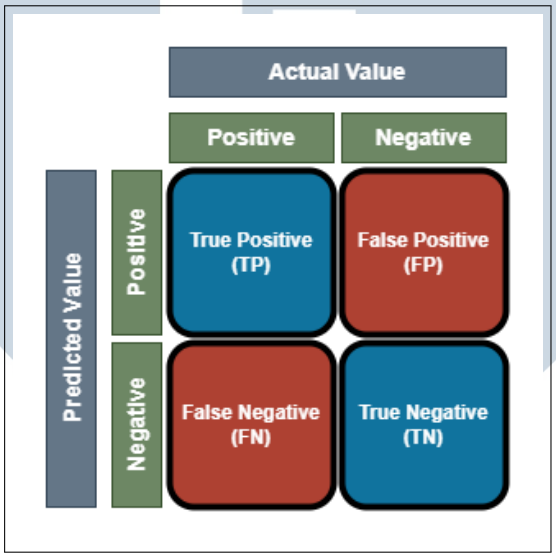
$$TF\text{-}IDF = TF_{(t,d)} \times IDF_{(t)} \tag{2.7}$$

Keterangan:

- t = Kata
- d = Dokumen
- $TF_{(t,d)}$ = Frekuensi jumlah t dalam d
- $IDF_{(t)}$ = Inverse Document Frequency t

2.12 *Confussion Matrix*

Sebuah *Confussion Matriks* Adalah model matriks yang memberikan perbandingan antara hasil klasifikasi jumlah prediksi yang benar dan salah berdasarkan nilai aktual yang diketahui[8]. Nilai murni positif (TruePostive), nilai murni negatif (FalseNegative), nilai murni negatif (FalsePositive), dan nilai murni negatif (TrueNegatif) ditampilkan oleh matriks ini untuk penyesuaian data berdasarkan hasil klasifikasi[36]. Berikut adalah model tabel Cofussion Matrix[8].



Gambar 2.2. Confussion Matrix Table

Keterangan:

- TP (True Positive): Jumlah data positif yang diklasifikasikan nilai benar (positif) yang diprediksi oleh model.
- FP (False Positive): Jumlah data positif yang diprediksikan salah diklasifikasikan sebagai nilai positif (diprediksi menjadi nilai negatif).
- FN (False Negative): Jumlah data negatif yang diprediksi salah diklasifikasikan sebagai ulasan negatif (prediksi menjadi nilai positif).
- TN (True Negative): Jumlah data negatif yang diklasifikasikan benar (negatif) oleh model.

Berdasarkan nilai-nilai diatas ini maka untuk mengevaluasi menggunakan matriks seperti Accuracy, Precision, Recall, dan F1-score, dibawah ini dapat dijabarkan dalam rumus berikut[36].

A. Accuracy

Accuracy adalah angka yang menunjukkan seberapa baik sebuah model bekerja secara keseluruhan pada semua kelas, Accuracy dihitung dengan cara membandingkan berapa banyak prediksi yang benar dengan total semua prediksi yang dibuat oleh model.

Rumus 2.8 menunjukkan cara perhitungan Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

B. Precision

Precision menunjukkan seberapa akurat model saat menetapkan sebuah data sebagai positif. Hasil ini melihat seberapa banyak prediksi positif yang benar.

Rumus 2.9 menunjukkan cara perhitungan Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.9}$$

C. Recall

Recall menilai dengan menemukan atau mengenali data yang benar-benar positif. Nilai ini dihitung dengan membandingkan jumlah data positif yang berhasil diprediksi dengan benar dengan total seluruh data yang sebenarnya positif.

Rumus 2.10 menunjukkan cara perhitungan Recall

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.10}$$

D. F1-Score

F1-Score Adalah nilai yang diperoleh dengan menghitung rata-rata harmonik dari precision dan recall. Hasil ini dihitung dengan menggabungkan precision dan recall menggunakan rumus rata-rata harmonik.

Rumus 2.11 menunjukkan cara perhitungan F1-Score

$$F1\text{-Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2.11}$$

2.13 *Text Preprocessing*

Proses *text preprocessing* bertujuan mengubah teks mentah menjadi bentuk yang lebih rapi dan mudah dipahami[13]. Didalamnya terdapat beberapa hal dilakukan seperti menghapus kata yang bukan bahasa Inggris, menghilangkan tanda baca, menghapus emotikon, memecah kalimat menjadi kata-kata kecil, serta menyamakan bentuk penulisan teks. Semua langkah ini membantu agar data lebih terstruktur, efisien, dan efektif untuk di analisis[8]. Berikut adalah langkah-langkah proses yang ada didalam Pre-processing[42]:

1. Cleaning Data

   Cleaning adalah proses memastikan kualitas data dengan cara menghapus atau memperbaiki bagian data yang tidak diperlukan, tidak lengkap, salah, atau memiliki format yang salah. Proses ini menghilangkan seperti URL, Tagar, Emotikon, username, maupen simbol-simbol.

2. Tokenization

   Tokenisasi adalah proses memecah teks atau komentar menjadi kata-kata terpisah. Pemisahan ini biasanya dilakukan berdasarkan spasi sehingga setiap kata bisa diproses secara lebih detail.

3. Case Folding

   Case folding proses menyeragamkan bentuk huruf dalam teks, biasanya dengan mengubah semua huruf besar yang ada (Uppercase) menjadi huruf kecil (Lowercase). Hal ini membantu sistem membaca data dengan lebih konsisten.

4. Normalization

   Normalization adalah proses pengoreksian kesalahan ejaan dan kata-kata yang tidak baku menjadi baku sesuai dengan aturan yang mengikuti Kamus Besar Bahasa Indonesia (KBBI)[16].

5. Stopword Removal

   Stopword removal adalah langkah menghapus kata-kata umum yang sering muncul tetapi tidak memberikan makna penting, seperti ”yang”, ”atau”, dan sejenisnya. Penghapusan ini membantu fokus pada kata-kata yang lebih informatif.

6. Stemming (Pencarian Kata Dasar)

   Stemming adalah proses mengubah kata berimbuhan menjadi bentuk dasar dengan menghilangkan awalan atau akhiran. Cara ini membantu menyamakan kata yang sebenarnya memiliki makna inti yang sama.

2.14 Text-blob

TextBlob menyediakan berbagai fitur untuk pemrosesan bahasa alami (NLP), seperti penandaan kelas kata (part-of-speech tagging), penerjemahan, analisis sentimen, klasifikasi, pengambilan kata benda maupun yang lainnya. Dalam analisis sentimen, TextBlob menghasilkan dua nilai yaitu polarity dan subjectivity[43]. Polarity menunjukkan seberapa positif atau negatif sebuah teks, dengan rentang -1 (negatif) hingga +1 (positif). Subjectivity menunjukkan seberapa subjektif atau objektif suatu teks, dengan nilai 0 sampai 1. Jika nilainya di bawah 0 maka teks lebih bersifat objektif daripada subjektif. Kalimat objektif bersifat fakta, sedangkan kalimat subjektif berisi perasaan pribadi, pandangan, keyakinan, opini, tuduhan, keinginan, asumsi, atau spekulasi.[44].

2.15 Easy Data Augmentation (EDA)

Easy Data Augmentation merupakan salah satu sebuah teknik *oversampling*, cara kerja nya dengan mengaugmentasikan data teks yang digunakan untuk mengatasi ketidakseimbangan data antar kelas dengan cara memperbanyak data pada kelas minoritas. Teknik ini memanfaatkan pendekatan sederhana seperti substitusi sinonim berdasarkan kamus bahasa Indonesia untuk menghasilkan variasi

data baru tanpa mengubah makna kalimat asli. Penerapan EDA membantu untuk dapat menghasilkan distribusi data yang lebih seimbang[45].

Easy Data Augmentation (EDA) merupakan salah satu teknik peningkatan data teks yang termasuk dalam pendekatan parafrase dan digunakan untuk mengatasi permasalahan ketidakseimbangan data. Metode ini bertujuan menambah jumlah data latih dengan menghasilkan variasi kalimat baru yang tetap mempertahankan makna dan konteks aslinya. EDA menerapkan empat teknik utama, yaitu *Synonym Replacement, Random Insertion, Random Swap,* dan *Random Deletion* yang digunakan untuk menciptakan keragaman pada data teks, Penerapan EDA juga terbukti mampu meningkatkan kinerja model pada tugas klasifikasi teks[46].