ANALISIS SENTIMEN TERHADAP SKEMA STUDENT LOAN UNTUK BIAYA PERGURUAN TINGGI PADA TWITTER MENGGUNAKAN ALGORITMA NAÏVE BAYES

Naufal Arib Fadhlurrohman, Aji Primajaya, Akbar Nugraha Dimyati
Informatika, Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat, Indonesia
naufalarfadh@gmail.com

ABSTRAK

Penelitian ini menganalisis sentimen masyarakat terhadap wacana penerapan skema student loan untuk biaya pendidikan tinggi di Indonesia. Data dikumpulkan dari media sosial Twitter atau X menggunakan Twitter API, menghasilkan 1.284 tweet setelah melalui tahap preprocessing. Metode yang digunakan adalah Knowledge Discovery in Database (KDD), meliputi data selection, preprocessing, transformation, data mining, dan evaluation. Pembobotan sentimen dilakukan menggunakan lexicon dan kamus kata negatif, kemudian dilanjutkan dengan metode TF-IDF. Hasil analisis menunjukkan terdapat 497 tweet positif, 501 negatif, dan 286 netral. Dataset dibagi menjadi data training dan testing dengan empat rasio pembagian, yaitu 90:10, 80:20, 70:30, dan 60:40. Model klasifikasi dibangun menggunakan algoritma Naïve Bayes Classifier, dengan pengujian performa melalui Confusion Matrix dan ROC AUC untuk mengukur akurasi, presisi, recall, F1-score, dan AUC. Penelitian ini memberikan manfaat teoritis berupa pengklasifikasian sentimen masyarakat terhadap skema student loan, serta manfaat praktis untuk pemerintah dalam memahami opini publik sebagai bahan pertimbangan kebijakan. Hasilnya menunjukkan bahwa analisis sentimen adalah pendekatan efektif untuk mengevaluasi tanggapan masyarakat terhadap kebijakan pendidikan, mendukung pengambilan keputusan yang lebih matang dalam penerapan skema student loan.

Kata kunci : analisis sentimen, *student loan, Naïve Bayes Classifier*, Twitter, *KDD*

1. PENDAHULUAN

Kini kita telah menduduki era digitalisasi yang sudah menjadi bagian dari hidup kita. Tentu ini membuat banyak perubahan dalam tata cara perilaku untuk manusia, salah satunya perubahan yang kita alami adalah bagaimana cara kita berdiskusi. Media sosial adalah tempat yang sangat mudah dijangkau untuk menyampaikan opini maupun pandangan terhadap berbagai isu, isu sosial, isu politik, isu nasional maupun isu global, termasuk isu pendidikan yang penulis akan bahas dalam penelitian ini**.**

Akhir-akhir ini permasalahan biaya perguruan tinggi tengah menjadi perbincangan yang hangat lantaran kampus Institut Teknologi Bandung (ITB) yang membuat pernyataan bahwasannya jika terkendala dengan biaya perguruan tinggi diarahkan untuk pinjaman uang *online*. Didalam artikel "Viral, Kasus Pinjol Bayar Kuliah yang Berujung pada Klarifikasi Danacita" yang diterbitkan oleh media Bisnis, pihak Bisnis menghubungi Yogi selaku Ketua Kabinet Keluarga Mahasiswa Institut Teknologi Bandung dan beliau mengungkapkan bahwasannya Rektor tiba-tiba membuat sebuah kebijakan. Kebijakannya yaitu jika tidak bisa membayar tunggakan uang kuliah tunggal (UKT) harus cuti. Ditengah kekisruhan yang terjadi, kemudian diberikanlah solusinya, salah satunya adalah Danacita yang kemudian di unggah dalam website pembayaran uang kuliah tunggal (UKT) Institut Teknologi Bandung (ITB) [27].

Wacana penerapan student loan di Indonesia mencuat setelah fenomena penggunaan pinjaman daring (pinjol) untuk pembayaran UKT di ITB dan pernyataan Menteri Keuangan Sri Mulyani yang membahas pengembangannya di LPDP [5].

Hal ini memicu diskusi publik, terutama di media sosial Twitter atau X, yang menjadi wadah populer untuk berbagi opini dan informasi. Penelitian ini memanfaatkan Twitter atau X sebagai sumber data untuk menganalisis sentimen masyarakat terhadap pengkajian skema student loan. Dengan lebih dari 500 tweet yang membahas topik ini, diperlukan metode untuk mengolah data tersebut.

Analisis sentimen, atau opinion mining, dipilih sebagai metode penelitian. Analisis sentimen merupakan cabang ilmu data mining yang bertujuan menganalisis, memahami, mengolah, dan mengekstrak opini dari data tekstual terhadap berbagai entitas [14].

Metode ini relevan dengan tujuan penelitian, yaitu mengklasifikasikan pandangan masyarakat terhadap student loan menjadi sentimen positif dan negatif. Hasil analisis sentimen ini diharapkan dapat memberikan gambaran mengenai penerimaan masyarakat terhadap wacana student loan.

Kebutuhan akan persiapan yang matang dalam penerapan student loan semakin mendesak mengingat tingginya penyaluran dana pinjol kepada mahasiswa, mencapai Rp450 miliar, dengan 83,6% di antaranya berasal dari Danacita, mitra ITB [6].

Opini publik yang dikumpulkan melalui analisis sentimen di Twitter atau X dapat menjadi bahan

pertimbangan penting bagi pemerintah, khususnya Menteri Keuangan, dalam merumuskan kebijakan terkait student loan. Dengan demikian, penelitian ini bertujuan menyediakan data dan informasi yang relevan untuk meminimalisir potensi permasalahan dan kekisruhan dalam implementasi skema student loan di Indonesia.

2. TINJAUAN PUSTAKA

Kandungan dalam tinjauan pustaka bisa berupa landasan teori yang relevan dengan topik skripsi yang akan dibahas, di mana landasan teori ini menjadi panduan pertama dalam penyusunan skripsi [3].

2.1. Media Sosial

Media sosial, sebagai arena jaringan global, memfasilitasi interaksi antar pengguna di seluruh dunia, menyederhanakan komunikasi jarak jauh secara signifikan [8].

Dengan pertumbuhan pengguna yang pesat dan penggunaan global, media sosial telah menjadi teknologi yang sangat berpengaruh dalam kehidupan sehari-hari.

Selain sebagai sumber informasi penting, media sosial juga memfasilitasi pengungkapan perasaan, opini, dan sentimen secara terbuka [2].

Kemampuan menghubungkan individu dan komunitas luas menjadikan media sosial bagian integral dari kehidupan modern [4], menciptakan ruang digital untuk interaksi dan pertukaran informasi tanpa batasan geografis. Singkatnya, media sosial telah merevolusi komunikasi dan interaksi jarak jauh serta memperluas akses informasi dan diskusi isu global.

2.2. Twitter / X

Twitter, yang kini dikenal sebagai X, merupakan platform media sosial populer, terutama di kalangan anak muda, yang memungkinkan pengguna berbagi informasi dan opini melalui *tweet*. Dengan pengalaman lebih dari sepuluh tahun, Twitter memiliki 528 juta pengguna aktif bulanan yang menghasilkan sekitar 237 juta *tweet* setiap hari [21].

Jumlah pengguna aktif yang besar, termasuk di Indonesia, menjadikan Twitter sebagai platform yang potensial untuk melakukan survei opini. Agar dapat dimengerti dengan baik berikut adalah statistik yang didapatkan dari Demandsage:

Gambar 1. Statistik pengguna twitter

2.3. *Student Loan*

Student loan, atau pinjaman mahasiswa, adalah skema pinjaman untuk membiayai pendidikan tinggi, terkadang termasuk biaya hidup, dengan pembayaran setelah lulus atau bekerja melalui cicilan sesuai gaji [19].

Skema ini pertama kali diterapkan di University of Bolognia pada akhir abad ke-11, diikuti universitas lain, dan diformalkan secara administrasi di Oxford University pada 1240. Kolombia menjadi negara pertama yang menerapkannya secara nasional pada 1951[1].

Tujuannya adalah meningkatkan akses pendidikan tinggi bagi masyarakat.

Di Indonesia, pernah ada Kredit Mahasiswa Indonesia (KMI) pada masa Orde Baru, dengan pembayaran setelah lulus dan jaminan ijazah [26].

Namun, KMI dinilai gagal karena rendahnya apresiasi korporasi terhadap ijazah, mengakibatkan banyak kredit terhenti. Pengalaman ini memberikan pelajaran penting tentang tantangan *student loan*, termasuk beban utang, risiko gagal bayar, dan pentingnya desain kebijakan yang tepat.

2.4. *Data Mining*

Data Mining adalah metode identifikasi pola dan tren bermanfaat dalam data berukuran besar [12], bertujuan menciptakan model/struktur baru yang praktis, optimal, dan mudah dipahami melalui proses iteratif dan kolaboratif. Proses ini melibatkan eksplorasi database besar untuk menemukan informasi yang mendukung pengambilan keputusan di masa depan, menggunakan alat analisis data mendalam untuk investigasi yang lebih detail [22].

2.5. Analisis Sentimen

Analisis sentimen adalah teknik untuk mengevaluasi dan mengategorikan emosi dalam teks menjadi positif atau negatif [3].

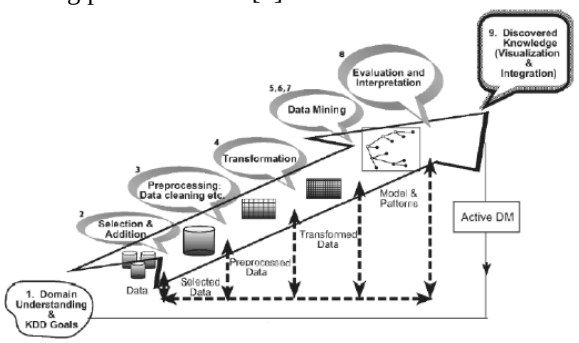
Metode ini penting karena emosi dan perspektif pembaca memengaruhi pemahaman dan respons terhadap suatu materi. Analisis sentimen berperan penting dalam mengklasifikasikan opini dalam teks, mengidentifikasi apakah opini tersebut positif atau negatif.

Menurut Chakraborty et al. [4], analisis sentimen adalah proses pengumpulan dan evaluasi perspektif individu terhadap kejadian sehari-hari. Dalam konteks media sosial, analisis sentimen menjadi teknologi untuk menangkap dan memproses opini di platform jejaring sosial. Semakin banyak data teks yang terkumpul, semakin mudah menemukan hubungan antar teks dan mengidentifikasi jenis sentimen, yang dapat dimanfaatkan untuk memperkirakan arah sentimen dalam berbagai teks.

2.6. KDD (*Knowledge Discovery in Database*)

Knowledge Discovery of Database (KDD) adalah pendekatan sistematis untuk menemukan pola yang andal, baru, dapat diterapkan, dan dapat dipahami

dalam kumpulan data dari berbagai ukuran atau kompleksitas. Berikut adalah gambaran singkat tentang prosedur KDD [7]:



Gambar 2. Tahapan *knowledge discovery in database*

Pengumpulan data akan diseleksi terlebih dahulu untuk melakukan tahap KDD selanjutnya, data hasil seleksi yang telah di proses di kumpulkan kembali dalam satu kumpulan berkas.

2.7. *Preprocessing*

Preprocessing adalah tahap krusial dalam *data mining* untuk membersihkan dan memperbaiki data mentah, mempersiapkannya agar sesuai kriteria metode analisis, dan menghasilkan analisis yang akurat dan andal.

2.8. *Transformation*

Tahap transformation mengubah struktur data agar sesuai dengan kriteria masing-masing metode dalam tahapan data mining.

2.9. *Data Mining*

Data mining, tahapan utama KDD, menghasilkan informasi dari dataset/dokumen melalui metode/algoritma seperti klasifikasi dan clustering.

2.10. *Evaluation*

Tahap ini mengevaluasi performa hasil penelitian, misalnya dengan menguji akurasi proses data mining menggunakan algoritma/metode tertentu.

2.11. *Crawling Data*

Crawling data adalah proses pengambilan dan pengindeksan cepat halaman web dalam jumlah besar dari repositori lokal untuk menanggapi permintaan kata kunci [13].

Informasi yang digunakan mesin pencari web berasal langsung dari situs web. Unggahan di Twitter, yang disebut "Tweet", dapat dibaca oleh pengguna lain dengan batasan 140 karakter [11].

Crawling data di Twitter melibatkan akses informasi pengguna dan tweet dari server Twitter melalui Twitter API (API) [20].

Meskipun awalnya ditujukan untuk koneksi dan pertukaran pengguna, Twitter API kini lebih umum digunakan untuk mempelajari komunitas dan opini tentang topik populer [17].

2.12. *Naïve Bayes Classifier*

Pengklasifikasi Naïve Bayes didasarkan pada teorema Bayes, menggunakan probabilitas bersyarat untuk setiap kelas potensial dari matriks data. Asumsi dasarnya adalah semua variabel beroperasi independen satu sama lain, meskipun asumsi ini mungkin tidak selalu benar dalam situasi dunia nyata. Namun, pengklasifikasi berbasis NBC tetap memberikan hasil memuaskan dalam berbagai masalah klasifikasi (Moraes et al., 2019). Algoritma Naïve Bayes dapat memperkirakan kemungkinan seseorang menjadi bagian dari suatu kelas dengan baik [9]. Teorema Bayes diilustrasikan dalam rumus berikut [25].

$$P(H|X) = \frac{p(X|H).p(H)}{p(X)}$$

Keterangan

X : Sampel data belum diketahui labelnya
H : Hipotesis bahwa x merupakan data label
P(H) : Peluang dari hipotesa H
P(X) : Peluang data sampel
P(X|H) : Peluang data sampel apabila hipotesis benar
P(H|X) : Peluang hipotesis berdasarkan keadaan sampel

Ketika X adalah pembuktian, H adalah hipotesis, P (H | X) adalah probabilitas bahwa hipotesis H benar untuk pembuktian X atau dengan kata lain P (H | X) merupakan probabilitas posterior H dengan syarat X. P (X| H) merupakan probabilitas posterior X dengan syarat H. P (H) adalah probabilitas prior hipotesis H, dan P (X) adalah probabilitas prior bukti X. Alur kerja dari Naive Bayes adalah sebagai berikut:

a. Mempelajari data pelatihan.
b. Menentukan probabilitas dengan menghitung jumlah kemungkinan hasil yang diberikan setiap faktor.
c. Temukan rata-rata, simpangan baku, dan probabilitas dalam sebuah tabel.

2.13. Contoh *Penerapan Naïve Bayes*

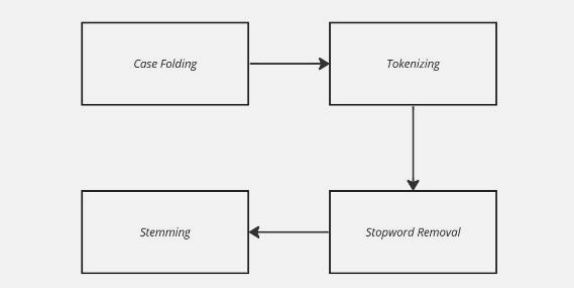
Naïve Bayes sering digunakan untuk klasifikasi teks, misalnya membedakan email spam dan bukan spam. Prosesnya meliputi:

a. Pengumpulan Data Pelatihan: Mengumpulkan email berlabel (spam/bukan spam) dan mengubahnya menjadi vektor fitur (misalnya, kata-kata dalam email).
b. Pembentukan Model: Menghitung probabilitas awal setiap kelas (spam/bukan spam) dan probabilitas kemunculan setiap kata di kedua kelas berdasarkan frekuensi kemunculan kata.
c. Klasifikasi Data Baru: Menghitung probabilitas kelas spam dan bukan spam untuk email baru menggunakan model yang sudah dibuat, lalu menetapkan kelas dengan probabilitas tertinggi sebagai kategori email tersebut.

2.14. *Text Mining*

Text mining adalah penemuan pengetahuan baru dari teks [23].

Teks adalah rangkaian kata dalam bahasa natural dengan berbagai arti yang digabungkan menggunakan aturan sintaks [28].



Gambar 3. Tahapan *text mining*

Tahapan *text mining* meliputi:

a. Case Folding: Mengubah semua huruf menjadi huruf kecil untuk standarisasi (misalnya, "DATA" menjadi "data"). Hanya karakter a-z yang dipertahankan, sisanya dihapus sebagai pemisah.
b. Tokenizing: Memecah teks menjadi kata-kata terpisah (token), misalnya "HUJAN DI KAMPUS UNSIKA" menjadi "HUJAN", "DI", "KAMPUS", "UNSIKA".
c. Stopword Removal: Menghapus kata-kata umum yang tidak signifikan (stopword) seperti "yang", "dan", "di", misalnya "Saya adalah mahasiswa di unsika yang terkenal indah" menjadi "Saya mahasiswa unsika terkenal indah".
d. Stemming: Mengurangi kata ke bentuk dasarnya (akar kata) untuk memperkecil jumlah indeks dan mengumpulkan kata dengan akar yang sama, misalnya "bersama", "kebersamaan" menjadi "sama".

2.15. *Confusion Matrix*

Confusion matrix adalah tabel m x m yang menggambarkan kinerja pengklasifikasi dengan membandingkan label prediksi dan label aktual [9]. Pengklasifikasi yang baik memiliki nilai di diagonal utama (*True Positives* dan *True Negatives*) dan nilai nol atau mendekati nol di luar diagonal. Berikut representasi *confusion matrix* 2x2:

Keterangan:
P: Jumlah data positif aktual
N: Jumlah data negatif aktual
P': Jumlah data positif prediksi
N': Jumlah data negatif prediksi
TP (*True Positives*): Prediksi positif benar
TN (*True Negatives*): Prediksi negatif benar
FP (*False Positives*): Prediksi positif salah (seharusnya negatif)
FN (*False Negatives*): Prediksi negatif salah (seharusnya positif)

Tiga metrik evaluasi penting yang dihitung dari *confusion matrix* adalah:

a. Accuracy: Seberapa baik klasifikasi secara keseluruhan.
   Accuracy = (TP + TN) / (TP + TN + FP + FN)
b. Precision: Seberapa tepat prediksi positif.
   Precision = TP / (TP + FP)
c. Recall: Seberapa baik model menemukan semua data positif yang sebenarnya.
   Recall = TP / (TP + FN)

Nilai *accuracy*, *precision*, dan *recall* biasanya dinyatakan dalam persentase (0-100%). Sistem dianggap baik jika ketiga metrik ini bernilai tinggi.

2.16. *Term Frequence-Inverse Document Frequence* (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang memberikan nilai numerik pada setiap kata atau fitur dalam teks, mengukur pentingnya kata dalam konteks teks tersebut. TF (Term Frequency) mengukur frekuensi kemunculan kata dalam dokumen; kata yang lebih sering muncul dalam satu dokumen memiliki bobot lebih tinggi, sementara kata yang sering muncul di banyak dokumen memiliki bobot lebih rendah [15].

2.17. *Python*

Python diakui sebagai bahasa pemrograman serbaguna teratas, yang kompatibel dengan berbagai sistem operasi modern. Selain itu, *Python* adalah bahasa pemrograman bertipe *interpreted language*, yang berarti kode program tidak dikonversi menjadi kode yang dapat dimengerti oleh komputer hingga saat dijalankan, dengan proses konversi ini berlangsung saat *runtime* [24].

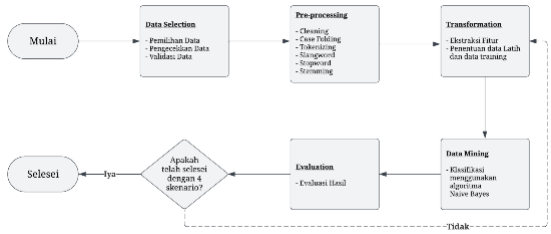
Python sendiri diciptakan oleh Guido Van Rossum. Aplikasi yang bisa dikembangkan dari aplikasi ini beragam salah satunya adalah di bidang *Scientific* dan numeric yang memiliki berbagai *library Scipy*, *Pandas*, dan *Scikit-learn* [24].

2.18. *Google Colaboratory*

Google Colaboratory (Google Colab) adalah layanan cloud yang memungkinkan pembuatan, eksekusi, dan pembagian kode Python langsung dari browser web. Ditujukan untuk analis, developer, peneliti, dan pendidik di bidang Data Science dan Machine Learning, Colab menyediakan lingkungan komputasi gratis yang mudah diakses dan fleksibel. Keunggulannya meliputi kemampuan menjalankan Jupyter Notebook tanpa konfigurasi, kolaborasi real-time seperti Google Docs, penyimpanan notebook di Google Drive untuk aksesibilitas dari mana saja, dan integrasi dengan library Python populer seperti TensorFlow, PyTorch, dan OpenCV untuk pengembangan Machine Learning yang lebih cepat dan efektif [18].

3. METODE PENELITIAN

Metodologi penelitian yang digunakan adalah *Knowledge Discovery in Database* (KDD), dimana tahapannya yaitu *Data Selection*, *Preprocesssing*, *Transformation*, *Data Mining*, *Evaluation*.



Gambar 4. Alur metodologi penelitian

3.1. *Data Selection*

Penelitian ini diawali dengan pemilihan data melalui *crawling* menggunakan Twitter API untuk mengumpulkan informasi terkait skema *student loan* dari platform X (dahulu Twitter). Pencarian dilakukan menggunakan tagar "skema student loan", "mahasiswa bayar pinjol", dan "pinjaman mahasiswa" untuk menggali tweet yang relevan. Setelah data terkumpul, tahap selanjutnya adalah penyaringan data untuk memastikan kesesuaiannya dengan kebutuhan dan kriteria penelitian.

3.2. *Preprocesssing*

Setelah seleksi data tweet, langkah selanjutnya adalah *preprocessing* untuk membersihkan data. Proses ini meliputi beberapa tahapan: *Cleaning*, yaitu penghapusan elemen non-esensial seperti URL, hashtag, username, dan simbol; *Case Folding*, yaitu konversi semua huruf menjadi huruf kecil; *Tokenizing*, yaitu pemisahan teks menjadi kata-kata berdasarkan spasi dan karakter non-alfanumerik, serta penghapusan karakter tunggal non-alfanumerik; *Slangword*, yaitu normalisasi kata-kata tidak baku menjadi bentuk baku; *Stopword Removal*, yaitu penghapusan kata-kata umum yang dianggap tidak informatif; dan *Stemming*, yaitu pengubahan kata berimbuhan ke bentuk dasarnya sesuai KBBI, contohnya "memberikan" menjadi "beri".

3.3. *Transformation*

Pada tahap transformasi, dilakukan pemilihan fitur dengan mengubah kata menjadi bentuk numerik secara manual melalui reduksi dan pemetaan data untuk mengidentifikasi fitur yang relevan untuk *data mining*. Selanjutnya, data dibagi menjadi dua segmen: *Data Training* yang digunakan untuk melatih model dengan data berlabel, dan *Data Testing* yang digunakan untuk memprediksi kategori/label data. Penelitian ini menggunakan empat skenario pembagian data *training* dan *testing*: 90%-10%, 80%-20%, 70%-30%, dan 60%-40%.

3.4. *Data Mining*

Pada tahap ini, data tweet diolah untuk diklasifikasikan sentimennya ke dalam tiga kategori: positif, negatif, dan netral. Klasifikasi ini dilakukan menggunakan algoritma Naive Bayes Classifier. Data yang sebelumnya telah dibagi menjadi data training dan data testing akan dianalisis untuk mendapatkan nilai-nilai metrik evaluasi seperti Akurasi (*Accuracy*), Presisi (*Precision*), *Recall*, F-measure, *true positive*, *true negative*, *false positive*, dan *false negative*. Singkatnya, tahap ini menggunakan Naive Bayes untuk mengkategorikan sentimen tweet dan mengukur performa klasifikasi tersebut.

3.5. *Evaluation*

Tahap evaluasi, sebagai bagian akhir KDD, menyimpulkan temuan dari *data mining* dan mengamati hasil prediksi sistem. Ketepatan prediksi diukur dengan *Confusion Matrix* untuk menghitung *Accuracy*, *Precision*, *Recall*, dan *F-measure* guna mengukur efektivitas hasil.

4. HASIL DAN PEMBAHASAN

Sesuai dengan tujuan penelitian untuk menganalisis sentimen data tweet mengenai skema *student loan* di Indonesia, bab ini akan memaparkan hasil yang diperoleh dan membahasnya secara mendalam.

4.1. Hasil Penelitian

Penelitian ini menganalisis sentimen tweet tentang skema *student loan* di Indonesia dan mengklasifikasikannya ke dalam tiga kelas: positif, negatif, dan netral. Klasifikasi dilakukan dengan algoritma *Naive Bayes Classifier*. Kinerja model dievaluasi menggunakan *Confusion Matrix* untuk mengukur akurasi (*accuracy*), presisi (*precision*), *recall*, dan F-measure.

4.2. *Data Selection*

Penelitian ini mengumpulkan data tweet dari Twitter menggunakan kata kunci "skema student loan", "pinjaman mahasiswa", dan "mahasiswa bayar pinjol" berbahasa Indonesia melalui Twitter API. Data awal yang terkumpul sebanyak 1444 tweet, terdiri dari 234 tweet dengan kata kunci "skema student loan", 690 tweet dengan "pinjaman mahasiswa", dan 520 tweet dengan "mahasiswa bayar pinjol". Data yang diperoleh masih belum terstruktur dan mengandung banyak noise seperti tanda baca, angka, simbol, dan kata tidak baku yang perlu dibersihkan sebelum proses klasifikasi. Data ini akan diproses lebih lanjut pada tahap Pre-Processing.

4.3. *Preprocesssing*

Tahap pre-processing data teks merupakan langkah krusial untuk membersihkan dan mempersiapkan data agar analisis sentimen lebih efektif. Data teks mentah sering mengandung noise yang dapat memengaruhi akurasi model. Tahapan ini

meliputi cleaning, yaitu menghapus elemen tidak relevan seperti URL, angka (kecuali dalam konteks tertentu), tanda baca, simbol (@, #, $, %, dll.), mention (@username), simbol hashtag (namun kata setelahnya dipertahankan), dan emoticon yang dikelompokkan menjadi kategori senang dan sedih. Tahap case folding dilakukan dengan mengubah semua huruf menjadi kecil untuk menyeragamkan data, sedangkan tokenizing memecah kalimat menjadi kata per kata, dan filtering menghapus kata-kata tidak penting (stopwords) seperti "jadi", "sih", "sudah", dan sebagainya. Selanjutnya, pada tahap stemming, kata diubah ke bentuk dasar dengan menghilangkan imbuhan (prefiks, sufiks, konfiks) menggunakan library Sastrawi, misalnya "berjalan" menjadi "jalan" atau "memasak" menjadi "masak". Tahap ini mengurangi variasi kata dan meningkatkan akurasi analisis. Setelah seluruh proses pre-processing selesai, jumlah dataset berkurang dari 1444 menjadi 1284, karena data yang tidak relevan untuk klasifikasi dihapus.

4.4. *Transformation Data*

Setelah melakukan tahap *Pre-Processing*, tahap selanjutnya adalah *transformation* data. Sebelum *transformation* data, dilakukan proses pembobotan sentimen untuk melabelkan data. Pada proses ini, skor sentimen dihitung dengan menjumlahkan kata-kata bersentimen positif dan negatif. Skor s7.6 cmentimen diperoleh dari jumlah polaritas sentimen positif dan negatif dalam dataset. Kamus *Lexicon* yang digunakan dalam penelitian ini adalah kamus *Lexicon* dan *Negative Word* dari penelitian sebelumnya oleh (Koto & Rahmaningtyas, 2017). Kamus ini digunakan untuk membobot ulang sentimen dari setiap tweet dan menilai sentimen secara lebih akurat oleh komputer.

Setelah menambahkan kamus *Lexicon* dan *negative words*, langkah berikutnya adalah menghitung skor sentimen dengan mengukur polaritas setiap kalimat dalam data tweet tersebut. Hasil pembobotan sentimen menggunakan kamus lexicon dan negative words terlihat bahwa setiap kata dalam kalimat diberi skor, kemudian semua skor tersebut dijumlahkan. Jika jumlah skor sentimen lebih dari 0, kalimat tersebut masuk ke dalam kelas positif. Jika skor sentimen kurang dari 0, kalimat tersebut masuk ke dalam kelas negatif. Jika nilai sentimen 0, kalimat tersebut masuk ke dalam kelas netral. Rincian jumlah data yang memiliki label positif, negatif, dan netral dari hasil pembobotan sentimen ini dapat dilihat pada Tabel 1.

Tabel 1. Jumlah Pada Setiap Sel

Label/Kelas	Jumlah
Positif	497
Negatif	501
Netral	286
Total	**1284**

Hasil dari pembobotan sentimen menggunakan kamus lexicon dan negative words berhasil mengelompokkan 1284 data tweet ke dalam 3 kelas, yaitu 497 data ke dalam kelas positif, 501 data ke dalam kelas negatif, dan 286 data ke dalam kelas netral. Selanjutnya, setelah pembobotan sentimen dengan kamus *lexicon* dan *negative words*, akan dilakukan pembobotan kata dengan mengimplementasikan TF-IDF (*Term Frequency – Inverse Document Frequency*). Fungsi yang digunakan adalah TfidfVectorizer yang terdapat pada library sklearn. Hasil dari proses pembobotan TF-IDF yang diterapkan diperlihatkan bahwa kata-kata/istilah yang terdapat dalam korpus diurutkan berdasarkan abjad, kemudian dihitung kemungkinan kemunculan kata-kata tersebut dalam suatu dokumen.

4.5. *Data Mining*

Setelah tahap *transformation data* selesai, langkah berikutnya adalah proses *data mining*. Tahap ini melibatkan klasifikasi data dengan mengimplementasikan salah satu model dari algoritma *Naïve Bayes*, yaitu *Multinomial Naïve Bayes*, menggunakan library `sklearn.naive_bayes` pada *Python*. Dalam penelitian ini, proses pengolahan data menggunakan algoritma *Naïve Bayes* dengan 4 skenario yang terdiri dari 90:10, 80:20, 70:30, dan 60:40. Hasil analisis sentimen tentang Skema *Student Loan* Untuk Biaya Perguruan Tinggi menggunakan *Naïve Bayes classifier* sebagai algoritma untuk pengklasifikasian, dan *Confusion Matrix* digunakan untuk menghasilkan nilai akurasi prediksi mesin terhadap data yang sudah diproses. Berikut adalah pembagian data training dan data testing.

Tabel 2. Tabel Skenario Pembagian *Split Data*

Presentase Data		Jumlah Data	
Training	**Testing**	**Training**	**Testing**
90%	10%	1163	129
80%	20%	10	257
70%	30%	898	386
60%	40%	770	514
Total		**1284**	

4.6. *Evaluation*

Setelah model selesai dibuat, model dari setiap skenario akan diuji. Nilai yang diperoleh dari pengujian meliputi *accuracy*, *precision*, *recall*, dan *f-measure (f1-score)*. Berikut ini adalah hasil evaluasi dari setiap model skenario yang diuji.

a. Skenario 1 (90% : 10%)
   Hasil klasifikasi dengan menggunakan 90% data training dan 10% data testing. Berdasarkan Gambar 4.9, hasil klasifikasi dengan *menggunakan Naïve Bayes* pada rasio data 90:10 menunjukkan nilai *accuracy* sebesar 51%, *precision* 34%, recall 44%, dan *f-measure* 38%.
b. Skenario 2 (80% : 20%)
   Hasil klasifikasi dengan menggunakan 80% data training dan 20% data testing. Berdasarkan Gambar 4.10, hasil klasifikasi dengan menggunakan *Naïve Bayes* pada rasio data 80:20

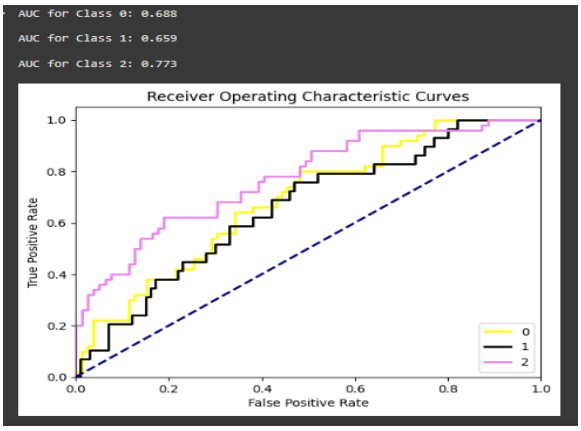
menunjukkan nilai *accuracy* sebesar 54%, *precision* 36%, recall 46%, dan *f-measure* 40%.

c. Skenario 3 (70% : 30%)
Hasil klasifikasi dengan menggunakan 70% data training dan 30% data testing. Hasil klasifikasi dengan menggunakan *Naïve Bayes pada* rasio data 70:30 menunjukkan nilai *accuracy* sebesar 52%, *precision* 35%, recall 45%, dan *f-measure* 39%.

d. Skenario 4 (60% : 40%)
Hasil klasifikasi dengan menggunakan 60% data training dan 40% data testing. Berdasarkan Gambar 4.12, hasil klasifikasi dengan menggunakan *Naïve Bayes* pada rasio data 60:40 menunjukkan nilai *accuracy* sebesar 51%, *precision* 34%, recall 44%, dan f-measure 38%. Selain menggunakan *confusion matrix* untuk mengukur performa model, AUC (*Area Under Curve*) dan kurva ROC (*Receiver Operating Characteristics*) juga dapat digunakan sebagai metode evaluasi. Hasil penerapan tahap AUC dapat dilihat pada Tabel 3.

Tabel 3. Hasil Nilai AUC Setiap Skenario

Skenario	AUC	Kualitas
90:10	0,71	*Good Classification*
80:20	0,66	*Fair Classification*
70:30	0,64	*Fair Classification*
60:40	0,63	*Fair Classification*



Gambar 5. Grafik ROC Hasil Skenario 90:10

Dari Tabel 3, hampir semua model skenario menunjukkan nilai AUC dengan kualitas *fair classification*. Nilai AUC terendah ditemukan pada model skenario 60:40 sebesar 0.63, sedangkan nilai AUC tertinggi terdapat pada model skenario 90:10 sebesar 0.71 dengan kualitas *good classification*. Berikut adalah grafik ROC dari model dengan nilai AUC tertinggi, yaitu pada skenario 90:10.

Pada Gambar 5 ditunjukkan nilai AUC untuk kelas positif dengan garis berwarna kuning sebesar 0.69, kelas netral dengan garis berwarna hitam sebesar 0.66, dan kelas negatif dengan garis berwarna merah muda sebesar 0.77. Berdasarkan nilai AUC dari tiap kelas tersebut, ROC AUC score pada model adalah:

$$\text{ROC AUC } Score = \frac{0{,}69+0{,}66+0{,}77}{3} = 0{,}71$$

Hal ini dapat diartikan bahwa model skenario 90:10 dengan nilai AUC tertinggi, yaitu 0.761, memiliki kualitas *good classification*.

4.7. Pembahasan

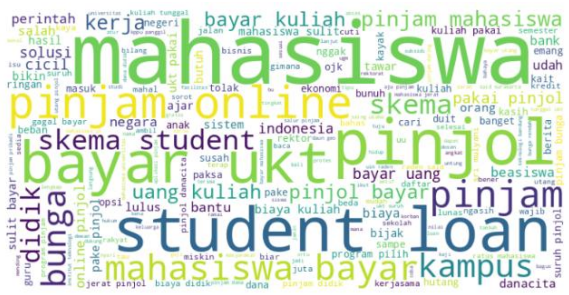
Penelitian ini menganalisis sentimen masyarakat terhadap skema *student loan* untuk biaya perguruan tinggi menggunakan metodologi Knowledge Discovery in Database (KDD). Data dikumpulkan melalui *crawling* dengan Twitter API menggunakan Python, menghasilkan 1.444 *tweet*. Setelah tahap *preprocessing* untuk membersihkan data dari noise, jumlah data berkurang menjadi 1.284 *tweet*.

Proses selanjutnya adalah pembobotan kata dengan TF-IDF dan pembentukan model klasifikasi menggunakan Multinomial Naive Bayes. Pengujian model dilakukan menggunakan *Confusion Matrix* dan ROC AUC untuk mengukur akurasi, presisi, *recall*, F1-score, serta AUC. Hasil pengujian keempat skenario dibandingkan untuk mengevaluasi performa model klasifikasi. Hasil perbandingan pengujian dari keempat model *Multinomial NB* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Perbandingan Evaluasi Model *Naive Bayes*

Model	*Accuracy*	*Precission*	*Recall*	*F. Measure*	*AUC*	Kualitas
90:10	51%	34%	44%	38%	0,71	*Good Classification*
80:20	54%	36%	46%	40%	0,66	*Fair Classification*
70:30	52%	35%	45%	39%	0,64	*Fair Classification*
60:40	51%	34%	44%	38%	0,63	*Fair Classification*

Tabel 4 menunjukkan perbandingan antar model skenario yang diuji. Untuk pengujian nilai *accuracy*, model 90:10 dan 60:40 memiliki nilai terendah dengan 51%, sedangkan nilai tertinggi terdapat pada model 80:20 dengan 54%. Pada *precision*, nilai terendah ada pada model 90:10 dan 60:40 dengan 34%, dan tertinggi pada 80:20 dengan 36%. Nilai *recall* tertinggi dimiliki oleh model 80:20 dengan 46%, sementara yang terendah ada pada model 90:10 dan 60:40 dengan 44%. *F-Measure* tertinggi terdapat pada model 80:20 dengan nilai 40%, dan terendah pada 90:10 dan 60:40 dengan 38%. Untuk nilai *AUC*, model 60:40 memiliki nilai terendah, sedangkan yang tertinggi ada pada model 90:10. Dari perbandingan skor pengujian keempat model skenario beserta nilai *AUC* yang didapat, model 90:10 dapat diartikan memiliki kualitas *good classification.*



Gambar 6. *Wordcloud keyword* 'mahasiswa bayar pinjol', 'pinjaman mahasiswa', dan *'skema student loan'*

Gambar 6 menunjukkan *Wordcloud* yang menampilkan kata-kata yang sering muncul dalam *dataset*. Kata-kata yang paling sering muncul adalah "mahasiswa", "pinjol", "*student loan*", "pinjam *online*", dan "bayar ukt". Semakin besar ukuran kata dalam *Wordcloud*, semakin tinggi frekuensi kemunculannya, yang menunjukkan bahwa kata tersebut sering digunakan oleh masyarakat dalam cuitan di Twitter. Berdasarkan hasil sentimen, kelas sentimen netral memiliki jumlah data yang jauh lebih banyak dibandingkan kelas sentimen lainnya. Hal ini menunjukkan bahwa Skema S*tudent Loan* Untuk Bbiaya Perguruan Tinggi di Indonesia cenderung masuk dalam kategori sentimen Negatif.

5. KESIMPULAN DAN SARAN

Penelitian analisis sentimen Skema Student Loan untuk Perguruan Tinggi menggunakan algoritma Naïve Bayes menghasilkan temuan sebagai berikut: Pengujian akurasi dengan metode Split Data (90:10, 80:20, 70:30, 60:40) menunjukkan model 80:20 memiliki akurasi dan presisi tertinggi, sedangkan nilai AUC tertinggi (0.71) terdapat pada model 90:10, menunjukkan kualitas good classification. Analisis sentimen Twitter pada tiga kelas (Positif, Netral, Negatif) menunjukkan 501 label Negatif, 497 Positif, dan 286 Netral, dengan sentimen Negatif mendominasi, menandakan kecenderungan masyarakat terhadap skema ini negatif..

Berdasarkan hasil penelitian ini, penelitian selanjutnya dapat dikembangkan lebih lanjut dalam aspek model klasifikasi. Peneliti menyarankan: Dataset yang digunakan pada penelitian selanjutnya sebaiknya lebih besar untuk meningkatkan akurasi data. Penelitian selanjutnya disarankan menggunakan dataset yang diperoleh dari media sosial lain. Penelitian selanjutnya dapat menggunakan fitur seleksi dengan metode *information gain* untuk meningkatkan akurasi.

DAFTAR PUSTAKA

[1] N. Aisyah, "Sejarah Student Loan di Dunia Berawal dari Kampus Ini," *Detik Media*, Jan. 12, 2022. [Online]. Available: https://www.detik.com/edu/detikpedia/d-5893953/sejarah-student-loan-di-dunia-berawal-dari-kampus-ini

[2] Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, "Sentiment analysis of extremism in social media from textual information," *Telematics and Informatics*, vol. 48, p. 101345, 2020. [Online]. Available: https://doi.org/10.1016/J.TELE.2020.101345

[3] H. Badjrie, O. N. Pratiwi, and H. D. Anggana, "Analisis Sentimen Review Customer terhadap Produk IndiHome dan First Media Menggunakan Algoritma Convolutional Neural Network," *E-Proceeding of Engineering*, vol. 8, no. 5, pp. 9049, 2021.

[4] Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Applied Soft Computing Journal*, vol. 97, p. 106754, 2020. [Online]. Available: https://doi.org/10.1016/j.asoc.2020.106754

[5] *CNN Indonesia*, "Mengenal Skema Student Loan yang Sedang Dikaji Sri Mulyani," Feb. 1, 2024. [Online]. Available: https://www.cnnindonesia.com/ekonomi/20240201085757-532-1057026/mengenal-skema-student-loan-yang-sedang-dikaji-sri-mulyani

[6] *CNN Indonesia*, "Utang Pinjol Rp450 M Sudah 'Menjerat' Mahasiswa," Feb. 26, 2024. [Online]. Available: https://www.cnnindonesia.com/ekonomi/20240226081128-78-1067204/utang-pinjol-rp450-m-sudah-menjerat-mahasiswa

[7] W. Gata and Purnomo, "Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS," *Format*, vol. 6, no. 1, pp. 1–13, 2017.

[8] G. Appel, L. Grewal, R. Hadi, and A. T. Stephen, "The future of social media in marketing," 2019.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. ScienceDirect, 2012. [Online]. Available: https://doi.org/10.1016/C2009-0-61819-5

[10] F. Koto and G. Rahmaningtyas, "InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," 2017. [Online]. Available: https://doi.org/10.1109/IALP.2017.8300625

[11] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," *Proc. Int. AAAI Conf. Web and Social Media*, vol. 5, no. 1, pp. 538–541, 2011. [Online]. Available: https://doi.org/10.1609/icwsm.v5i1.14185

[12] D. Larose and C. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed. 2014. [Online]. Available: https://doi.org/10.1002/9781118874059

[13] B. Liu, *Web Data Mining*, 2nd ed., vol. 1. Springer, 2011.

[14] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.

[15] R. Melita, V. Amrizal, H. B. Suseno, and T. Dirjam, "Penerapan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan Cosine Similarity pada Sistem Temu Kembali Informasi untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018. [Online]. Available: http://dx.doi.org/10.15408/jti.v11i2.8623

[16] R. Moraes, E. Soares, and L. Machado, "A double weighted fuzzy gamma naive bayes classifier," *J. Intell. & Fuzzy Syst.*, vol. 38, pp. 1–12, 2019. [Online]. Available: https://doi.org/10.3233/JIFS-179431

[17] H. Nguyen and R. Zheng, "A Data-driven Study of Influences in Twitter Communities," *Proc. IEEE Int. Conf. Commun.*, pp. 1–6, 2014. [Online]. Available: https://doi.org/10.1109/ICC.2014.6883936

[18] *REVOU*, "Apa itu Google Colab? Revolusi Cita Edukasi," 2024. [Online]. Available: https://revou.co/kosakata/google-colab

[19] R. Sebayang, "Mengenal Apa Itu Student Loan yang Mirip KTA Perbankan," *CNBC Indonesia*, Feb. 16, 2019. [Online]. Available: https://www.cnbcindonesia.com/tech/20190216185334-37-55929/mengenalapa-itu-student-loan-yang-mirip-kta-perbankan

[20] J. E. Sembodo, E. Setiawan, and A. Baizal, "Data Crawling Otomatis pada Twitter," pp. 11–16, 2016. [Online]. Available: https://doi.org/10.21108/INDOSC.2016.111

[21] R. Shewale, "Twitter Statistics For 2024 — (Facts After 'X' Rebranding)," *Demandsage*, Jan. 10, 2024. [Online]. Available: https://www.demandsage.com/twitter-statistics

[22] E. D. Sikumbang, "Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori," *J. Tek. Komput.*, vol. 4, no. 1, 2018.

[23] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and Semantic Issues of Text Mining," *ACM Sigmod Record*, vol. 36, no. 3, pp. 23–34, 2007.

[24] A. Stewart, *Python Programming: Python Programming for Beginners, Python Programming for Intermediates, Python Programming for Advanced*. CreateSpace Independent Publishing Platform, 2017.

[25] Suntoro, *Data Mining: Algoritma dan Implementasi dengan Pemrograman PHP*. Elex Media Komputindo, 2019.

[26] L. Ulfa H, "Student Loan as a Funding Solution for College Student in Indonesia," Mar. 2015.

[27] P. H. Untari, "Viral Kasus Pinjol Bayar Kuliah yang Berujung pada Klarifikasi Danacita," *Bisnis Media*, Feb. 3, 2024. [Online]. Available: https://finansial.bisnis.com/read/20240203/563/1737922/viral

[28] J. Žižka, F. Dařena, and A. Svoboda, *Text Mining with Machine Learning*. Boca Raton, 2019. [Online]. Available: https://doi.org/10.1201/9780429469275