

Analisis Sentimen *Twitter* Tentang Pinjaman *Online* di Indonesia Menggunakan Metode *Random Forest*

Muhammad Alif A.^{1*}, Muslim Alamsyah², Mochammad Firman Arif³
 Teknik Informatika, Universitas Merdeka Pasuruan, Pasuruan, Indonesia
 *e-mail Corresponding Author: muhammadalifardiansyah19@gmail.com

Abstract

The rapid growth of digital technology has led to a significant surge in online lending in Indonesia. However, inadequate regulations and unethical practices by service providers have generated diverse public opinions, particularly on social media platforms such as Twitter. This study aims to analyze public sentiment towards online lending by applying the Random Forest algorithm to Twitter data. Random Forest was chosen for its ability to handle overfitting and provide accurate classification results. The dataset consisted of 1,000 tweets categorized into positive, negative, and neutral sentiments. Before classification, data preprocessing was conducted, including text cleaning, tokenization, stemming, and TF-IDF calculation. Results demonstrated that the combination of Random Forest and the SMOTE technique significantly contributed to sentiment analysis. The developed model achieved an accuracy of 80%, with precision scores of 83%, 65%, and 83% for negative, neutral, and positive sentiments, respectively. Recall scores were 93%, 55%, and 45% for negative, neutral, and positive sentiments, respectively, while F1-scores were 88%, 59%, and 59%. In conclusion, although the model exhibited optimal performance, especially for negative sentiment, further improvements are needed for a more comprehensive analysis, particularly for neutral sentiment.

Keywords: *Online lending; Random forest; Sentiment analysis, SMOTE; Twitter.*

Abstrak

Di tengah kemajuan pesat era digital, sektor pinjaman online di Indonesia mengalami pertumbuhan signifikan. Namun, regulasi yang belum memadai dan meningkatnya tindakan tidak etis oleh penyedia layanan telah menimbulkan beragam pandangan di kalangan masyarakat, terutama pada platform media sosial seperti Twitter. Tujuan dari penelitian ini adalah menganalisis persepsi publik tentang pinjaman online dengan menerapkan metode *Random Forest* pada teks *tweet* di Twitter. Metode ini dipilih karena kemampuannya dalam mengatasi *overfitting* dan memberikan hasil klasifikasi yang akurat. Data penelitian terdiri dari 1000 *tweet* yang dikategorikan sebagai sentimen positif, negatif, atau netral. Sebelum klasifikasi, dilakukan *preprocessing* data termasuk pembersihan teks, tokenisasi, dan stemming serta dilakukan perhitungan TF-IDF. Hasil penelitian menunjukkan bahwa kombinasi metode *Random Forest* dan teknik SMOTE memberikan kontribusi signifikan dalam analisis sentimen. Model yang dikembangkan mencapai akurasi 80%, dengan *precision* untuk sentimen negatif 83%, netral 65%, dan positif 83%. Nilai *recall* untuk sentimen negatif adalah 93%, netral 55%, dan positif 45%, sedangkan *F1-score* masing-masing adalah 88%, 59%, dan 59%. Kesimpulan dapat disimpulkan bahwa meskipun model ini menunjukkan kinerja yang optimal, terutama untuk sentimen negatif, peningkatan lebih lanjut diperlukan untuk analisis yang lebih komprehensif, terutama pada sentimen netral.

Kata kunci: *Analisis sentimen; pinjaman online; Random forest; SMOTE; Twitter.*

1. Pendahuluan

Seiring dengan pesatnya perkembangan era digital, teknologi kini telah menjadi elemen yang sangat penting dalam kehidupan sehari-hari, memengaruhi cara kita berinteraksi, bekerja, dan berkomunikasi. Revolusi industri 4.0 telah membawa inovasi besar di berbagai sektor, termasuk sektor keuangan, yang menghasilkan model bisnis baru dan mendisrupsi perbankan tradisional. Salah satu inovasi signifikan dalam sektor ini adalah *Financial Technology (Fintech)*[1], yang menghadirkan layanan keuangan lebih efisien dan aman melalui proses

transaksi yang modern[2]. Di antara berbagai layanan yang ditawarkan oleh *Fintech*, pinjaman online atau dikenal sebagai "pinjol" telah menjadi salah satu solusi populer bagi masyarakat yang membutuhkan akses kredit cepat dan mudah[3].

Di Indonesia, sektor pinjaman online telah menunjukkan perkembangan yang pesat, didorong oleh kemudahan akses dan proses administrasi yang cepat [4]. Berdasarkan data Otoritas Jasa Keuangan (OJK), industri *fintech lending* mencatatkan peningkatan laba bersih jumlahnya mencapai Rp 50,48 miliar pada bulan Januari 2023, dibandingkan dengan kerugian yang dialami pada tahun sebelumnya. Meskipun kemudahan akses dan transaksi tanpa memerlukan agunan khusus pada pinjaman online memberikan keuntungan, pertumbuhan ini juga menghadapi berbagai tantangan. Tantangan tersebut meliputi regulasi yang belum memadai, serta meningkatnya kasus penipuan dan perilaku yang tidak sesuai dengan etika oleh penyedia layanan pinjaman online [5]. Hal ini menyebabkan masyarakat sering kali terjebak dalam godaan pinjaman online, yang apabila tidak dibayar tepat waktu, dapat terjerat bunga yang sangat tinggi, dan jika diabaikan peminjam bisa menghadapi sanksi hukum atas dasar penipuan terkait pembayaran pinjaman yang telah disepakati[6]. Hal ini telah memunculkan berbagai pandangan dan opini di kalangan masyarakat, yang banyak disuarakan melalui *platform* media sosial, salah satunya adalah Twitter[7].

Twitter, yang kini dikenal sebagai "X" kerap menjadi media bagi pengguna untuk berbagi pendapat dan pengalaman mengenai berbagai isu terkini, mulai dari hiburan, politik, sosial, hingga pemerintahan. Istilah 'pinjol' sering muncul dalam percakapan tersebut, mencerminkan beragamnya sentimen masyarakat terhadap layanan ini[8]. Dalam upaya untuk memahami sentimen ini, dilakukan analisis sentimen, yaitu metode untuk mengambil dan menentukan pendapat atau data subjektif dari teks terkait suatu masalah atau peristiwa, yang dikelompokkan menjadi kategori netral, positif, atau negatif. Proses ini melibatkan teknik komputasi untuk memahami sentimen yang terkandung dalam teks. Studi ini memanfaatkan metode *Natural Language Processing* (NLP), ini merupakan teknik canggih dalam komputasi bahasa alami untuk menganalisis dan merepresentasikan teks seperti bahasa manusia[9]. Sebelum penerapan algoritma, dilakukan tahapan *preprocessing data* yang mencakup pembersihan data, penghapusan karakter yang bukan huruf, dan penghilangan istilah yang tidak relevan. Proses ini memastikan penggunaan data telah dioptimalkan untuk pengklasifikasian[10]. Tahapan *preprocessing* ini terdiri dari beberapa tahapan penting seperti *cleaning*, *case folding*, *stopword removal*, *tokenization*, dan *stemming*.

Untuk memahami sentimen masyarakat terhadap pinjaman online dengan lebih baik, Tujuan dari penelitian ini adalah mengidentifikasi dan mengkategorikan sentimen public di Indonesia mengenai pinjaman online menjadi tiga jenis kategori, yaitu negatif, netral, dan positif, melalui analisis data tweet. Selain itu, penelitian ini juga mengevaluasi efektivitas metode *Random Forest* dalam menganalisis sentimen tweet mengenai pinjaman online di Indonesia. Dipilih Metode *Random Forest* karena kemampuannya dalam menangani *overfitting* dan memberikan hasil klasifikasi yang akurat dengan menggunakan beberapa pohon keputusan.

2. Tinjauan Pustaka

Penelitian yang berjudul "Analisis Sentimen Pinjaman Online Di Twitter Menggunakan Algoritma *Support Vector Machine* (SVM)"[11]. Menganalisis sentimen ulasan masyarakat tentang pinjaman online dengan mengimplementasi metode *Support Vector Machine* (SVM), untuk mengklasifikasikan pandangan masyarakat mengenai pinjaman online di *platform* Twitter menggunakan *software Rapid Miner Studio* yang menghasilkan nilai akurasi untuk pinjaman online sebesar 62%. Klasifikasi menunjukkan sebesar 59% sentimen negatif dan 41% sentimen Positif.

Analisa serupa dengan judul "Analisis Sentimen Pinjaman Online Di Media Sosial *Twitter* Menggunakan Metode *Naive Bayes*" [8] dengan 2931 dataset awal yang diambil dari *Twitter* menggunakan aplikasi Rapidminer. Setelah melalui proses *text preprocessing* terdapat Jumlah data yang tersedia adalah 2.912, yang selanjutnya memberikan pembobotan menggunakan metode TF-IDF. Dengan nilai hasil akhir sentimen negatif sebesar 68.61% dengan 1998 data, dan sentimen positif 31.39% dengan 914 data. Analisis sentimen tersebut menghasilkan tingkat akurasi sebesar 80%.

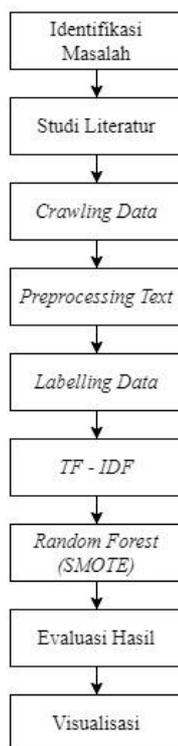
Penelitian lain "Perbandingan Metode K-NN Dan Metode *Random Forest* Untuk Analisis Sentimen pada *Tweet* Isu Minyak Goreng di Indonesia" dengan hasil pengujian dari metode K-

NN yaitu akurasi 72,86%, sedangkan hasil pada metode *Random Forest* yaitu akurasi 73,37%[12].

Selanjutnya penelitian lain yang memiliki judul “*Sentiment Analysis* Terkait Pemindahan Ibu Kota Indonesia Menggunakan Metode *Random Forest* Berdasarkan *Tweet* Warga Negara Indonesia. Penelitian ini bertujuan untuk menganalisis tanggapan terhadap pemindahan ibu kota Indonesia pasca pengesahan Undang-Undang IKN, dan juga membandingkan jumlah *tweet* yang memiliki sentimen positif, negatif, dan netral. Menggunakan metode *Random Forest* dengan memperoleh sentimen positif sebesar 55%, negatif 14%, dan netral 32% dari data total. Dengan hasil tingkat akurasi tertinggi yaitu mencapai 76%, dengan *recall* 70%, presisi 69% dan *f1-score* 69%[13].

Penelitian sebelumnya telah menunjukkan beragam hasil dalam analisis sentimen terkait pinjaman online dan topik lainnya, dengan penggunaan metode algoritma yang berbeda. Meskipun algoritma-algoritma tersebut memiliki keunggulan masing-masing namun, masalah ketidakseimbangan kelas dalam dataset sentimen sering kali menjadi tantangan utama, mengingat distribusi yang tidak merata antara kelas yang sering muncul dalam data sentimen. Mengingat hal tersebut, penelitian ini bertujuan untuk memperluas jumlah data yang digunakan dan menerapkan metode *Random Forest*, yang dikenal karena kemampuannya dalam mengatasi berbagai jenis ketidakseimbangan kelas. Lebih lanjut, penelitian ini mengintegrasikan teknik *Synthetic Minority Over-sampling Technique* (SMOTE), sebuah metode penyeimbangan data yang memperluas dataset minoritas dengan menciptakan contoh sintesis baru. Kombinasi *Random Forest* dan SMOTE ini merupakan pendekatan yang belum banyak diterapkan dalam konteks analisis sentimen, khususnya dalam topik pinjaman online di Indonesia.

3. Metodologi



Gambar 1. Alur Penelitian

Dalam penelitian ini, terdapat beberapa tahapan alur yang akan diterapkan. Tahapan-tahapan tersebut mencakup identifikasi masalah, diikuti dengan kajian literatur dan pengumpulan data yang dilakukan melalui proses *crawling* pada platform media sosial Twitter (X). Selanjutnya, data yang telah dikumpulkan akan melalui proses *preprocessing*, yang meliputi *case folding*, *remove punctuation*, *tokenize*, *stopword removal*, *stemming*. Setelah

preprocessing, data diberi label dan kata-kata dalam dataset diberi bobot menggunakan metode TF-IDF. Selanjutnya, data dianalisis menggunakan algoritma *Random Forest* untuk klasifikasi. Evaluasi hasil dilakukan dengan menggunakan *confusion matrix* dan hasilnya divisualisasikan dalam bentuk *Word Cloud*. Gambar 1 menunjukkan alur penelitian secara keseluruhan.

3.1. Identifikasi Masalah

Penulis melakukan *research experiment* bagaimana sentimen masyarakat Indonesia terhadap pinjaman online dapat diidentifikasi dan dikategorikan menjadi sentimen positif, negatif, atau netral melalui analisis data tweet. Penelitian ini juga bertujuan untuk mengevaluasi efektivitas metode *Random Forest* dalam menangani ketidakseimbangan data yang sering terjadi dalam analisis sentimen. Penulis menerapkan teknik-teknik tersebut untuk menganalisis sentimen tweet mengenai pinjaman online di Indonesia.

3.2. Studi Literatur

Dalam penelitian ini, dilakukan eksplorasi literatur dengan mencari referensi dari buku, jurnal, dan situs internet yang relevan dengan topik dan judul penelitian.

3.3. Crawling Data

Data diperoleh melalui teknik *crawling* yang dilakukan dengan *Google Colab* dan bahasa pemrograman *Python* untuk mengidentifikasi sentimen mengenai pinjaman online di Indonesia. Dalam proses ini, sebanyak 1000 *tweet* diambil dari platform Twitter (X) yang diperoleh secara bertahap sejak bulan Januari, menggunakan kata kunci seperti "#pinjamanonline", "#pinjol", dan "#fintech".

3.4. Preprocessing Text

Pada tahap ini, data *tweet* diolah menjadi bentuk yang lebih terstruktur agar dapat memberikan informasi berkualitas yang telah siap untuk diterapkan dalam tahap selanjutnya [14]. Langkah-langkah tersebut meliputi pembersihan data dari *noise*, normalisasi teks, serta penghapusan unsur-unsur yang tidak relevan. Terdapat beberapa langkah *preprocessing* data yang diterapkan dalam penelitian ini:

- 1) *Case Folding*
Case Folding merupakan tahapan mengubah seluruh huruf dalam data tweet menjadi huruf kecil (*lowercase*). Dengan tujuan untuk menyelaraskan format teks dan memudahkan analisis.
- 2) *Remove Punctuation*
Remove Punctuation merupakan tahapan pengolahan teks yang melibatkan penghapusan tanda baca dari suatu teks seperti hashtag, koma, titik, tanda seru, dan lainnya dihapus dari teks *tweet*. Penghapusan ini dilakukan untuk menyederhanakan representasi teks dan mempermudah analisis lebih lanjut.
- 3) *Tokenize*
Tokenisasi merupakan proses memecah teks dalam *tweet* menjadi bagian-bagian lebih kecil seperti kata, frasa, atau elemen teks lainnya.
- 4) *Stopword Removal*
Stopword Removal merupakan proses makna signifikan, seperti "dan," "atau," "di," dan "adalah." Langkah ini memfokuskan analisis pada kata-kata yang memiliki makna lebih signifikan.
- 5) *Stemming*
Stemming merupakan tahapan menghapus akhiran atau awalan dari kata dalam teks *tweet*, sehingga hanya menyisakan bentuk dasar. Ini bertujuan untuk menyederhanakan teks dan menyatukan variasi kata yang memiliki arti yang serupa. Proses ini dilakukan dengan bantuan *library Sastrawi*.

3.5. Labelling Data

Pelabelan data dilakukan secara manual dengan menetapkan label positif, netral, atau negatif pada masing-masing sentimen.

3.6. TF-IDF

Langkah berikutnya setelah proses pelabelan data adalah penerapan metode *Term Frequency - Inverse Document Frequency* (TF-IDF). Metode ini diterapkan untuk mengukur frekuensi kemunculan sebuah kata dalam dataset serta frekuensi kemunculannya secara keseluruhan. Frekuensi kemunculan kata dalam dataset menggambarkan seberapa sering kata tersebut digunakan[15]. Rumus untuk menghitung nilai TF-IDF terlihat pada Persamaan 1 dan 2.

$$Wt = tf_t \times idf_t \quad (1)$$

$$idf_t = \log \frac{N}{df_t} \quad (2)$$

Dimana tf merujuk frekuensi kemunculan suatu kata dalam setiap tweet, N adalah total tweet, dan df_t adalah jumlah total tweet yang mengandung kata t .

3.7. Random Forest

Random Forest adalah metode yang menerapkan pendekatan yang menggunakan teknik *ensemble learning* (kumpulan pohon keputusan) dalam klasifikasi atau regresi[16]. Metode ini membangun banyak pohon keputusan secara acak, menggunakan subset acak dari data pelatihan serta fitur-fitur yang ada, guna mengurangi kemungkinan *overfitting* pada data. Proses pembuatan pohon keputusan melibatkan perhitungan *impurity Gini* dan *Gini gain* untuk menentukan pemisahan data yang paling optimal dapat dilihat pada persamaan 3 dan 4. Hasil dari setiap pohon keputusan kemudian digabungkan melalui mekanisme *voting* atau *averaging*[17].

$$Gini = 1 - \sum_{i=1}^a P_i^2 \quad (3)$$

$$Gini\ Gain = Gini - Weight\ Averaged\ Gini \quad (4)$$

Keterangan:

P_i = proporsi data yang berada dalam kelas i .

a = jumlah total kelas.

3.8. SMOTE

Synthetic Minority Over-sampling (SMOTE) merupakan teknik *oversampling* yang diterapkan untuk menangani masalah ketidakseimbangan dalam distribusi kelas suatu data[18]. Dalam analisis sentimen, ketidakseimbangan ini sering muncul ketika data sentimen positif memiliki jumlah yang jauh lebih besar atau lebih kecil dibandingkan dengan sentimen negatif atau netral. SMOTE mengatasi hal ini dengan menghasilkan data sintesis dari kelas minoritas guna menciptakan distribusi kelas yang lebih seimbang[19].

3.9. Evaluasi

Kinerja model yang telah dikembangkan dievaluasi melalui *confusion matrix*, yang memberikan beberapa metrik evaluasi seperti akurasi, *precision*, *recall*, dan *F1-score*.

3.9. Visualisasi

Visualisasi data menggunakan *Word Cloud*. *Word Cloud* adalah visualisasi frekuensi kemunculan kata dalam dokumen. Frekuensi kemunculan suatu kata dalam teks yang dianalisis semakin tinggi, maka kata tersebut akan tampil dengan ukuran yang lebih besar pada gambar *Word Cloud*[15]. Visualisasi memungkinkan identifikasi topik-topik utama yang sering dibahas oleh pengguna Twitter mengenai pinjaman online.

4. Hasil dan Pembahasan

4.1. Dataset

Data yang digunakan pada penelitian ini diperoleh dari media sosial twitter (X) melalui proses *crawling* yang diperoleh secara bertahap sejak bulan Januari, menggunakan kata kunci seperti “#pinjamanonline”, “#pinjol”, dan “#fintech”. Proses pengambilan data dilakukan menggunakan *Google Colab* dan bahasa pemrograman *Python*. Sebanyak 1000 *tweets* berhasil dikumpulkan dan disimpan dalam format CSV. Data hasil *crawling* terlihat pada tabel 1.

Tabel 1. Sampel Hasil *Crawling Data*

| No | Tweet |
|------|---|
| 1 | sobat jatim pasti sudah pernah mendengar mengenai layanan pinjaman online atau pinjol dong walaupun di keadaan sulit layanan ini seakan memberikan kemudahan dan jalan terang tetaplah cermat dan pahami risiko aplikasi pinjaman online ya |
| 2 | wakil ketua komisi dpr republik indonesia ahmad sahrani mendorong polri dan otoritas jasa keuangan otoritas jasa keuangan memberantas aplikasi pinjaman online pinjol ilegal penegak hukum juga diminta intensif memantau pinjol yang terdaftar di otoritas jasa keuangan |
| ... | ... |
| 999 | jika boleh jujur komentar ya perihal praktik pinjaman dan gagal bayar dari pinjaman online ini menurutku titik masalahnya bukan lagi pada konsumen tapi penyedia jasa hampir tiap saat kita di pengaruhi tentang pinjaman online bahkan dikasih akses mudah |
| 1000 | gue kira org di kejar pinjol tu cma lebay tp kemrn gue liat sendiri temen gue terlilit pinjol dia d maki oleh si penagih bahkan di ancam serem bgt pinjem 500 harus bayar 1 juta dlm sminggu trs d tf 1 jt lg pdhal dia gk pnjem pas mau balikin hrus balikin 2 jt |

4.2. Preprocessing Data

Prapemrosesan data merupakan langkah awal yang krusial dalam analisis sentimen ini. Data *tweet* yang memuat opini tentang pinjaman online di Indonesia telah melalui beberapa tahap prapemrosesan untuk menjamin data yang dihasilkan bersih serta siap digunakan oleh algoritma *Random Forest*. Tahapan ini mencakup pembersihan teks (*cleaning text*), *case folding*, *remove punctuation*, tokenisasi (*tokenize*), *stopword removal*, dan *stemming*. Dengan mengurangi *noise* dan mengekstrak fitur linguistik yang relevan, diharapkan model dapat lebih akurat dalam mengidentifikasi pola sentimen dalam data *tweet*. Langkah-langkah yang diterapkan dalam prapemrosesan data dalam penelitian ini meliputi:

4.2.1. Case Folding

Case folding merupakan teknik dalam pemrosesan teks yang bertujuan untuk mengubah seluruh huruf dalam data *tweet* menjadi huruf kecil (*lowercase*). Proses ini penting untuk memastikan konsistensi teks, sehingga analisis selanjutnya dapat dilakukan tanpa mempertimbangkan perbedaan antara huruf besar dan huruf kecil. Sebagai ilustrasi, penggunaan *case folding* pada data terlihat pada Tabel 2

Tabel 2. Proses *Case Folding*

| Sebelum | Sesudah |
|---|---|
| @Heraloebss Sistem pinjol seperti ini membuat resah masyarakat. Jalan praktis yang ditawarkan bertemu dengan desakan ekonomi yang terus membelenggu menjadi celah transaksi yang sangat tidakimbang untuk peminjam. Sistemnya harus segera dihapus. | @heraloebss sistem pinjol seperti ini membuat resah masyarakat. jalan praktis yang ditawarkan bertemu dengan desakan ekonomi yang terus membelenggu menjadi celah transaksi yang sangat tidakimbang untuk peminjam. sistemnya harus segera dihapus. |

4.2.2. Remove Punctuation

Remove Punctuation merupakan teknik pengolahan teks yang bertujuan untuk menghapus tanda baca dari teks seperti hashtag, titik, tanda seru, koma, dan simbol lainnya dihapus dari data *tweet* guna menyederhanakan representasi teks dan memfasilitasi proses analisis. Sebagai contoh penerapan dari penghapusan tanda baca ini, terlihat pada Tabel 3.

Tabel 3. Proses *Remove Punctuation*

| Sebelum | Sesudah |
|--|--|
| @heraloebss sistem pinjol seperti ini membuat resah masyarakat. jalan praktis yang ditawarkan bertemu dengan desakan ekonomi yang terus membelenggu menjadi celah transaksi yang sangat tidak imbang untuk peminjam. sistemnya harus segera dihapus. | sistem pinjol seperti ini membuat resah masyarakat. jalan praktis yang ditawarkan bertemu dengan desakan ekonomi yang terus membelenggu menjadi celah transaksi yang sangat tidak imbang untuk peminjam. sistemnya harus segera dihapus. |

4.2.3. Tokenize

Tokenisasi merupakan teknik dalam pengolahan teks yang bertujuan untuk membagi teks dari data *tweet* menjadi bagian-bagian yang lebih kecil seperti kalimat, kata, frasa, atau elemen teks lainnya, sesuai dengan metode tokenisasi yang digunakan. Contoh penerapan tokenisasi terlihat pada Tabel 4.

Tabel 4. Proses *Tokenize*

| Sebelum | Sesudah |
|--|---|
| sistem pinjol seperti ini membuat resah masyarakat. jalan praktis yang ditawarkan bertemu dengan desakan ekonomi yang terus membelenggu menjadi celah transaksi yang sangat tidak imbang untuk peminjam. sistemnya harus segera dihapus. | ['sistem', 'pinjol', 'seperti', 'ini', 'membuat', 'resah', 'masyarakat', 'jalan', 'praktis', 'yang', 'ditawarkan', 'bertemu', 'dengan', 'desakan', 'ekonomi', 'yang', 'terus', 'membelenggu', 'menjadi', 'celah', 'transaksi', 'yang', 'sangat', 'tidak', 'imbang', 'untuk', 'peminjam', 'sistemnya', 'harus', 'segera', 'dihapus'] |

4.2.4. Stopword Removal

Stopword Removal merupakan teknik pengolahan teks yang bertujuan untuk menghilangkan kata-kata umum yang tidak memiliki peran signifikan terhadap arti dari sebuah teks dalam data *tweet*. Kata-kata umum yang dihapus biasanya meliputi "dan," "atau," "di," "adalah," dan sebagainya. Penerapan penghilangan *Stopword* terlihat pada Tabel 5.

Tabel 5. Proses *Stopword Removal*

| Sebelum | Sesudah |
|---|--|
| sistem pinjol seperti ini membuat resah masyarakat jalan praktis yang ditawarkan bertemu dengan desakan ekonomi yang terus membelenggu menjadi celah transaksi yang sangat tidak imbang untuk peminjam sistemnya harus segera dihapus | sistem pinjol membuat resah masyarakat jalan praktis ditawarkan bertemu desakan ekonomi membelenggu celah transaksi tidak imbang peminjam sistemnya segera dihapus |

4.2.5. Stemming

Stemming merupakan teknik pengolahan teks yang bertujuan untuk menghilangkan akhiran atau awalan dari kata pada data *tweet*, sehingga hanya tersisa bentuk dasar kata tersebut tanpa imbuhan. Contoh penerapan *stemming* dapat dilihat pada Tabel 6.

Tabel 6. Proses *Stemming*

| Sebelum | Sesudah |
|--|---|
| sistem pinjol membuat resah masyarakat jalan praktis ditawarkan bertemu desakan ekonomi membelenggu celah transaksi tidak imbang peminjam sistemnya segera dihapus | sistem pinjol buat resah masyarakat jalan praktis tawar temu desakan ekonomi belenggu celah transaksi tidak imbang pinjam sistem segera hapus |

4.3. Pelabelan Data

Proses pelabelan data dalam penelitian ini dilakukan dengan teliti dan manual dan dibantu oleh seorang ahli dengan latar belakang pendidikan bahasa Indonesia. Setiap cuitan atau *tweet* yang berkaitan dengan pinjaman online di Indonesia dikategorikan ke dalam tiga kelas sentimen, yaitu positif, negatif, atau netral. Metode manual ini dipilih untuk memastikan tingkat akurasi dan konsistensi dalam penilaian sentimen, mengingat adanya variasi ekspresi sentimen di Twitter. Dataset yang telah dilabeli tersebut selanjutnya digunakan sebagai dasar untuk melatih model *Random Forest*. Hasil pelabelan terlihat pada tabel 7.

Tabel 7. Pelabelan Data

| No | Tweet | Label |
|------|---|---------|
| 1 | sobat jatim dengar layan pinjam online pinjol sulit layan akan mudah jalan terang tetap cermat paham risiko aplikasi pinjam online ya | positif |
| 2 | wakil ketua komisi dpr republik indonesia ahmad sahrani dorong polri otoritas jasa keuangan otoritas jasa keuangan berantas aplikasi pinjam online pinjaman online ilegal tegak hukum intensif pantau pinjaman online daftar otoritas jasa keuangan | netral |
| ... | ... | ... |
| 999 | jujur komentar ya perihal praktik pinjam gagal bayar pinjam online turut titik konsumen sedia jasa pengaruh pinjam online kasih akses mudah | negatif |
| 1000 | gue orang kejar pinjaman online lebay kemarin gue lihat teman gue lilit pinjaman online maki si tagih ancam seram banget pinjam bayar minggu transfer gak pinjam pas balikin | negatif |

4.4. TF-IDF

Pada tahap pra-pemrosesan data, *tweet* yang telah dibersihkan dari *noise* dan kemudian diproses menjadi bentuk dasar kata melalui *stemming*. Selain itu, setiap *tweet* diberi label untuk keperluan analisis. Selanjutnya, bobot kata dihitung menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Metode ini menghitung seberapa sering suatu kata muncul pada dokumen (*term frequency*) dan membandingkannya dengan frekuensi kemunculan kata-kata yang terdapat pada semua dokumen dalam korpus (*inverse document frequency*). Kosakata yang sering muncul dalam sebuah dokumen tetapi jarang ditemukan di dokumen lain akan memiliki bobot TF-IDF tinggi. Menandakan bahwa kata tersebut memiliki relevansi dan kekhususan yang besar terhadap dokumen tersebut. Pada tahap ini, peneliti menggunakan library *Scikit-learn* dengan *TfidfVectorizer* untuk menghitung bobot TF-IDF. Hasil dari perhitungan TF-IDF ditampilkan dalam tabel 8.

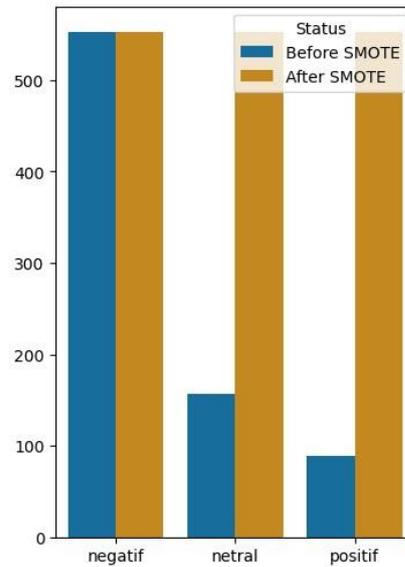
Tabel 7. Sampel Kata hasil TF-IDF

| Kata | Skor TF-IDF |
|--------|---------------------|
| pinjol | 0,644399781295525 |
| online | 0,428289233232437 |
| pinjam | 0,4241399222364545 |
| uang | 0,2722396975827112 |
| orang | 0,26950722034465503 |

4.5. Klasifikasi *Random Forest*

Dataset yang berisi teks *tweet* telah diproses dengan teknik praproses dan diwakili oleh fitur TF-IDF. Dataset ini kemudian dibagi secara acak ke dalam dua subset dengan rasio 90% untuk data latih dan 10% untuk data uji. Model klasifikasi *Random Forest* diterapkan untuk menentukan sentimen pada *tweet*. Untuk mengatasi ketidakseimbangan kelas yang sering muncul dalam data sentimen, teknik SMOTE digunakan. Model *Random Forest*, yang merupakan algoritma *ensemble* yang efektif, dipilih untuk klasifikasi sentimen. *Hyperparameter* model dioptimalkan guna mencapai performa terbaik. Evaluasi kinerja model dilakukan secara menyeluruh menggunakan *confusion matrix* untuk mengukur metrik seperti akurasi, presisi, recall, dan F1-score. Penggunaan *Pipeline* dari *imblearn* memfasilitasi integrasi yang efisien

dari seluruh proses, mulai dari praproses hingga evaluasi. Teknik SMOTE dapat dilihat pada gambar 2.



Gambar 2. Distribusi SMOTE

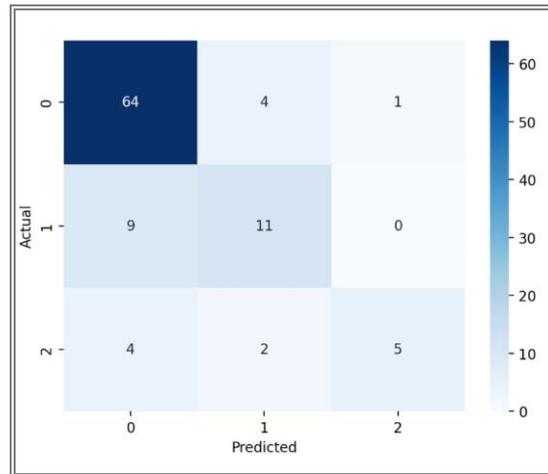
Setelah menerapkan teknik SMOTE untuk menyeimbangkan distribusi data, proses klasifikasi dilanjutkan menggunakan metode *random forest*. Hasil dari proses klasifikasi menunjukkan bahwa metode *random forest* mencapai tingkat akurasi sebesar 80%. Untuk metrik *precision*, nilai untuk label sentimen negatif adalah 83%, netral 65%, dan positif 83%. Sedangkan untuk nilai *recall*, label sentimen negatif adalah 93%, netral 55%, dan positif 45%. Adapun nilai *F1-score* masing-masing kategori adalah negatif 88%, netral 59%, dan positif 59%. Hasil dari proses klasifikasi menggunakan algoritma *random forest* ditampilkan pada gambar 3.

| Classification Report: | | | | |
|------------------------|-----------|--------|----------|---------|
| | precision | recall | f1-score | support |
| negatif | 0.83 | 0.93 | 0.88 | 69 |
| netral | 0.65 | 0.55 | 0.59 | 20 |
| positif | 0.83 | 0.45 | 0.59 | 11 |
| accuracy | | | 0.80 | 100 |
| macro avg | 0.77 | 0.64 | 0.69 | 100 |
| weighted avg | 0.79 | 0.80 | 0.79 | 100 |

Gambar 3. Hasil Klasifikasi *Random Forest*

4.6. Evaluasi

Kinerja model yang telah dikembangkan dievaluasi menggunakan *confusion matrix*. Alat ini berfungsi untuk menganalisis seberapa efektif hasil klasifikasi dan mengidentifikasi kesalahan antara kelas-kelas yang berbeda. *Confusion matrix* menyediakan berbagai metrik evaluasi, termasuk akurasi, *precision*, *recall*, dan *F1-score*. Metrik-metrik ini menyajikan informasi yang jelas mengenai kinerja model dalam melakukan klasifikasi data.



Gambar 4. Tabel Confusion Matrix

Berdasarkan tabel *confusion matrix* pada gambar 4, model berhasil mengklasifikasikan dengan tepat 64 sampel dari kelas 0 (negatif). Namun, terdapat 4 sampel yang salah diklasifikasikan sebagai kelas 1 (netral), dan 1 sampel yang salah dikategorikan sebagai kelas 2 (positif). Untuk kelas 1 (netral), model secara akurat mengklasifikasikan 11 sampel. Akan tetapi, 9 sampel dari kelas 1 (netral) salah dikategorikan sebagai kelas 0 (negatif), dan tidak ada sampel yang salah diklasifikasikan sebagai kelas 2 (positif). Sementara itu, untuk kelas 2 (positif), model berhasil mengklasifikasikan dengan benar 5 sampel. Terdapat 3 sampel yang salah dikategorikan sebagai kelas 0 (negatif), dan 2 sampel yang salah diklasifikasikan sebagai kelas 1 (netral).

4.7. Visualisasi

Untuk memperoleh pemahaman yang lebih mendalam tentang kata-kata yang sering muncul digunakan metode visualisasi berupa *Word Cloud*. Teknik ini memungkinkan untuk mengidentifikasi topik-topik utama yang sering dibahas dalam *tweet*.



Gambar 5. Visualisasi Word Cloud

Berdasarkan hasil visualisasi pada gambar 5, frekuensi kemunculan kata yaitu "pinjam online" "pinjam" "orang" "pinjol" "uang".

4.8. Pembahasan

Penelitian ini memperluas pengetahuan mengenai penggunaan metode Random Forest dalam analisis sentimen dengan fokus pada pinjaman online di Indonesia. Penelitian sebelumnya telah menunjukkan hasil yang bervariasi dalam aplikasi Random Forest untuk analisis sentimen pada berbagai konteks. Sebagai contoh, penelitian dengan judul "*Sentiment Analysis* Terkait Pemindahan Ibu Kota Indonesia Menggunakan Metode *Random Forest*

Berdasarkan Tweet Warga Negara Indonesia" menunjukkan bahwa *Random Forest* dapat digunakan secara efektif untuk analisis sentimen, dengan akurasi tertinggi mencapai 76% dan nilai recall, presisi, serta F1-score yang cukup baik, yaitu 70%, 69%, dan 69%. Meskipun demikian, penelitian ini tidak mengatasi secara mendalam masalah ketidakseimbangan kelas yang sering muncul dalam data sentimen[13]. Sementara itu, studi lain berjudul "Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest" mencatat akurasi yang sangat tinggi sebesar 94% dengan dataset yang jauh lebih besar, yaitu 3862 tweet, menegaskan konsistensi efektivitas Random Forest dalam analisis sentimen[20]. Penelitian tambahan dengan judul "Perbandingan Metode K-NN Dan Metode Random Forest Untuk Analisis Sentimen pada Tweet Isu Minyak Goreng di Indonesia" menunjukkan akurasi metode Random Forest sebesar 73,37%, sedikit lebih tinggi dibandingkan dengan metode K-NN yang mencapai 72,86% [12]. Penelitian ini memperkuat temuan-temuan sebelumnya dengan menunjukkan bahwa kombinasi Random Forest dan teknik SMOTE tidak hanya meningkatkan akurasi tetapi juga menangani masalah ketidakseimbangan kelas secara lebih efektif dalam konteks pinjaman online di Indonesia. Dengan demikian, penelitian ini memberikan wawasan baru yang relevan dan melengkapi literatur yang ada, serta menunjukkan bahwa pendekatan inovatif dalam penyeimbangan data dapat lebih meningkatkan performa metode Random Forest dalam klasifikasi sentimen.

5. Simpulan

Kesimpulan dari penelitian ini menunjukkan bahwa penerapan metode *Random Forest* bersama teknik SMOTE memberikan kontribusi signifikan dalam analisis sentimen terhadap pinjaman online di Indonesia, dengan data yang diambil dari Twitter. Dari dataset yang terdiri atas 1000 tweet, model yang dikembangkan mampu mencapai akurasi sebesar 80%. Hasil evaluasi lanjutan menunjukkan nilai presisi untuk label sentimen negatif sebesar 83%, sentimen netral 65%, dan sentimen positif 83%. Sementara itu, nilai *recall* untuk sentimen negatif mencapai 93%, sentimen netral 55%, dan sentimen positif 45%. Nilai F1-score yang diperoleh adalah 88% untuk sentimen negatif, 59% untuk sentimen netral, dan 59% untuk sentimen positif. Meskipun hasil ini cukup memuaskan, penelitian ini juga menunjukkan adanya peluang untuk pengembangan lebih lanjut, terutama dalam meningkatkan kinerja model pada kategori sentimen netral, yang dapat memberikan analisis yang lebih komprehensif di masa depan.

Daftar Referensi

- [1] H. S. Disemadi, "Fenomena *Predatory Lending*: Suatu Kajian Penyelenggaraan Bisnis *Fintech* P2P Lending selama Pandemi COVID-19 di Indonesia," *Pandecta Res. Law J.*, vol. 16, no. 1, pp. 55–67, 2021.
- [2] A. Ikhsan, M. F. A. Kusuma, A. C. M. Wibowo, and N. A. Rakhmawati, "Pengaruh Akun BOT pada Sentiment Masyarakat terhadap Pinjaman Online di Twitter".
- [3] U. Prajogo and R. Rusno, "Persepsi risiko terhadap minat melakukan pinjaman online dengan kemudahan penggunaan sebagai variabel moderasi," *MBR Manag. Bus. Rev.*, vol. 6, no. 1, pp. 22–32, 2022.
- [4] T. P. Lestari, "Analisis *Text Mining* pada Sosial Media Twitter Menggunakan Metode *Support Vector Machine* (SVM) dan *Social Network Analysis* (SNA)," *J. Inform. Ekon. Bisnis*, pp. 65–71, 2022.
- [5] R. Binekasri, "Kinerja Pinjol Makin Membaik, OJK Ungkap Penyebabnya," CNBC Indonesia. Accessed: Aug. 14, 2024. [Online]. Available: <https://www.cnbcindonesia.com/market/20230317130007-17-422572/kinerja-pinjol-makin-membaik-ojk-ungkap-penyebabnya>
- [6] R. Kartikawati and S. Soediro, "Kontroversi Pinjaman Online Ditinjau Dari Perspektif Moral," *Kosmik Huk.*, vol. 22, no. 3, pp. 246–267, 2022.
- [7] M. Apriyanto, "Dampak Konsumen Terhadap Pinjaman Online (PINJOL)," *PaKMas J. Pengabd. Kpd. Masy.* vol. 3, no. 1, pp. 52-58, 2023.
- [8] M. I. Ghazali, W. H. Sugiharto, and A. F. Iskandar, "Analisis Sentimen Pinjaman Online Di Media Sosial Twitter Menggunakan Metode *Naive Bayes*," *KLIK Kaji. Ilm. Inform. Dan Komput.*, vol. 3, no. 6, pp. 1340–1348, 2023.
- [9] D. A. Sani and M. Z. Sarwani, "Koreksi Jawaban Esai Berdasarkan Persamaan Makna Menggunakan *Fasttext* dan Algoritma *Backpropagation*," *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 11, no. 2, pp. 92–111, 2022.

- [10] J. Muliawan and E. Dazki, "Sentiment Analysis of Indonesia's Capital City Relocation Using Three Algorithms: Naïve Bayes, KNN, and Random Forest," *J. Tek. Inform. JUTIF*, vol. 4, no. 5, pp. 1227–1236, 2023.
- [11] D. S. Utami and A. Erfina, "Analisis Sentimen Pinjaman Online di Twitter Menggunakan Algoritma *Support Vector Machine* (SVM)," presented at the Prosiding Seminar Nasional Sistem Informasi dan Manajemen Informatika Universitas Nusa Putra, 2021, pp. 299–305.
- [12] C. P. Yanti, N. W. E. Agustini, N. L. W. S. R. Ginantra, and D. A. P. Wulandari, "Perbandingan Metode K-NN Dan Metode *Random Forest* Untuk Analisis Sentimen pada *Tweet* Isu Minyak Goreng di Indonesia," *J. Media Inform. Budidarma*, vol. 7, no. 2, pp. 756–765, 2023.
- [13] T. C. Herdiyani and A. U. Zailani, "Sentiment Analysis Terkait Pemindahan Ibu Kota Indonesia Menggunakan Metode *Random Forest* Berdasarkan *Tweet* Warga Negara Indonesia," *J. Teknol. Sist. Inf.*, vol. 3, no. 2, pp. 154–165, 2022.
- [14] S. Khairunnisa, A. Adiwijaya, and S. Al Faraby, "Pengaruh *Text Preprocessing* terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)," *J. Media Inform. Budidarma*, vol. 5, no. 2, pp. 406–414, 2021.
- [15] S. Sumayah, F. Sembiring, and W. Jatmiko, "Analysis of sentiment of Indonesian community on metaverse using support vector machine algorithm," *J. Tek. Inform. JUTIF*, vol. 4, no. 1, pp. 143–150, 2023.
- [16] R. Wijanarko, D. E. Ratnawati, and P. P. Adikara, "Analisis Sentimen Dampak Perkembangan *Artificial Intelligence* (AI) pada Media Sosial X/Twitter Menggunakan Metode *Random Forest*," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 8, no. 5, 2024.
- [17] T. F. Basar, D. E. Ratnawati, and I. Arwani, "Analisis Sentimen Pengguna Twitter terhadap Pembayaran *Cashless* menggunakan *Shopeepay* dengan Algoritma *Random Forest*," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 6, no. 3, pp. 1426–1433, 2022.
- [18] A. F. Anjani, D. Anggraeni, and I. M. Tirta, "Implementasi *Random Forest* Menggunakan SMOTE untuk Analisis Sentimen Ulasan Aplikasi *Sister for Students* UNEJ," *J. Nas. Teknol. Dan Sist. Inf.*, vol. 9, no. 2, pp. 163–172, 2023.
- [19] M. Y. Aldean, P. Paradise, and N. A. S. Nugraha, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode *Random Forest Classifier* (Studi Kasus: Vaksin *Sinovac*)," *J. Inform. Inf. Syst. Softw. Eng. Appl. INISTA*, vol. 4, no. 2, pp. 64–72, 2022.
- [20] O. Manullang, C. Prianto, and N. H. Harani, "Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan *Lexicon Based* Dan *Random Forest*," *J. Ilm. Inform.*, vol. 11, no. 02, pp. 159–169, 2023.