Pengelompokan Pengguna Gagal Bayar Pinjaman Online pada Media Sosial Twitter Menggunakan TF-IDF dan K-Means Clustering

Erfian Junianto[1], Vini Shiva Salshabila[*2]
Universitas Adhirajasa Reswara Sanjaya, Bandung, Jawa Barat, Indonesia 40282
e-mail: [1]erfian.ejn@ars.ac.id, [2]vinishivasalshabila06@gmail.com
*Penulis Korespondensi

Diterima: 13 Juli 2025; Direvisi: 8 November 2025; Disetujui: 18 November 2025

Abstrak

Maraknya layanan pinjaman daring di Indonesia telah mempermudah akses keuangan tetapi juga menyebabkan peningkatan kasus gagal bayar. Kejadian ini tidak dapat semata-mata dikaitkan dengan keterbatasan ekonomi pengguna, sehingga memerlukan eksplorasi yang lebih komprehensif tentang motif yang mendasarinya. Penelitian ini bertujuan untuk mengklasifikasikan motif di balik gagal bayar di kalangan pengguna pinjaman daring berdasarkan wacana publik di Twitter. Sebanyak 2.607 tweet berbahasa Indonesia dikumpulkan menggunakan metode crawling dengan token otorisasi dan alat tweet-harvest. Temuan tersebut mengungkapkan empat tema dominan: tekanan ekonomi dan perilaku keuangan berisiko, promosi yang menyesatkan, intimidasi penagih utang, dan gaya hidup konsumtif. Term Frequency–Inverse Document Frequency (TF-IDF) dan K-Means efektif dalam mengekstraksi informasi yang tidak terstruktur dan mengelompokkan opini publik. Studi ini berkontribusi pada pemahaman perilaku default dan mendukung pengembangan sistem deteksi risiko sosial berbasis teks di masa depan.

Kata kunci: Gagal Bayar, *K-Means Clustering,* Motif Sosial, Pinjaman *Online*, TF- IDF, Twitter

Abstract

The rise of online lending services in Indonesia has facilitated financial access but has also led to an increase in Failed to pay cases. This phenomenon cannot be solely attributed to users' economic limitations, thus requiring a more comprehensive exploration of the underlying motives. This study aims to classify the motives behind among online lending users based on public discourse on Twitter. A total of 2,607 Indonesian-language tweets were collected using a crawling method with authorization tokens and the tweet-harvest tool. The findings revealed four dominant themes: economic pressure and risky financial behavior, misleading promotions, debt collector intimidation, and a consumptive lifestyle. Term Frequency–Inverse Document Frequency (*TF-IDF) and K-Means were effective in extracting unstructured information and clustering public opinion. This study contributes to the understanding of default behavior and supports the development of future text-based social risk detection systems.*

Keywords*: Failed to pay, K-Means Clustering, Online Loans, Social Motives, TF-IDF, Twitter.*

1. PENDAHULUAN

Pertumbuhan layanan pinjaman daring di Indonesia meningkat secara signifikan seiring dengan perkembangan teknologi keuangan digital (*Financial Technology*). Kemudahan akses,

proses verifikasi minimal, serta pencairan dana yang cepat menarik minat pengguna lintas lapisan masyarakat. Namun demikian, kemudahan tersebut juga menimbulkan konsekuensi berupa peningkatan kasus gagal bayar, yang tidak semata disebabkan oleh keterbatasan finansial, melainkan juga oleh faktor perilaku dan sosial.

Dengan akses pinjaman *online* seperti ini menyebabkan permasalahan yang krusial yaitu meningkatnya kasus gagal bayar. Kondisi gagal bayar tidak hanya disebabkan oleh kurangnya kemampuan dalam mengelola keuangan, melainkan juga oleh berbagai motif penyalahgunaan seperti penggunaan praktik gesek tunai, penipuan berbasis identitas fiktif, serta penyebaran data pribadi oleh penyedia pinjaman [1], [2]. Permasalahan ini tidak hanya memberikan dampak Negative terhadap pengguna, tetapi juga mengakibatkan kerugian bagi penyedia layanan pinjaman *online*, karena beresiko gagal bayar yang tinggi berdampak pada perununan kepercayaan publik terhadap layanan keuangan digital [3]. Presepsi negatif terhadap layanan pinjaman *online* juga diperkuat oleh kurangnya pengawasan serta rendahnya literasi digital di kalangan Masyarakat [4].

Faktor lain yang turut berkontribusi adalah tekanan sosial, penipuan digital, rendahnya literasi keuangan, dan eksploitasi data pengguna. Menurut penelitian sebelumnya, beberapa pengguna gagal membayar bukan karena tidak mampu, tetapi karena promosi yang menyesatkan, pencurian identitas, atau campur tangan pihak ketiga seperti joki pinjol. Kondisi yang kompleks ini memunculkan pertanyaan penelitian yang penting: Apa motif dominan di balik gagal bayar pinjaman *online*, dan bagaimana hal ini dapat diidentifikasi menggunakan data media sosial.

Penelitian ini menggunakan teknik *Natural Language Processing* (NLP) dan algoritma clustering untuk mengeksplorasi perilaku gagal bayar dari konten yang dibuat pengguna di platform media X (sebelumnya Twitter). Menggunakan *Term Frequency–Inverse Document Frequency* (TF-IDF) dan *K-Means Clustering*, penelitian ini bertujuan untuk mengungkap pola dominan dan mengklasifikasikan opini publik ke dalam kelompok topik yang jelas terkait dengan gagal bayar pinjaman.

Meskipun telah terdapat beberapa penelitian yang memanfaatkan kombinasi TF-IDF dan K-Means untuk analisis teks di media sosial, sebagian besar studi tersebut hanya berfokus pada aspek teknis pengelompokan tanpa menelaah makna sosial dari hasil klaster. Kesenjangan inilah yang ingin diisi oleh penelitian ini, yaitu dengan menafsirkan hasil klaster secara tematik untuk mengungkap motif sosial dan perilaku gagal bayar yang mendasarinya, serta mengaitkannya dengan literatur perilaku keuangan dan sosial ekonomi pengguna pinjaman daring.

2. LANDASAN TEORI

Dalam penelitian ini, teori-teori yang dikaji meliputi konsep dasar tentang pinjaman *online* dan gagal bayar, pengantar *Data Mining, Machine Learning*, serta metode *Clustering* khususnya *K-Means.* Selain itu, dibahas pula representasi data teks menggunakan TF-IDF untuk mengolah data tidak berlabel. Kajian ini juga mengaitkan hasil-hasil penelitian sebelumnya yang memiliki kesamaan konteks maupun pendekatan, sebagai bahan perbandingan dan penguat argumentasi ilmiah dalam penelitian ini.

2.1. *Data Mining*

Data Mining merupakan proses penemuan pola tersembunyi dalam kumpulan data besar dengan tujuan ekstraksi informasi yang mungkin berguna dan dapat dipahami [5], menguraikan bahwa *Data Mining* mencakup langkah-langkah: persiapan data, pemodelan, evaluasi, dan interpretasi hasil, serta digunakan di berbagai domain seperti perbankan, pemasaran, serta kesehatan. Adapun tujuan utamanya adalah mengubah data mentah menjadi pengetahuan, yang menjadi dasar dalam pengambilan keputusan strategis.

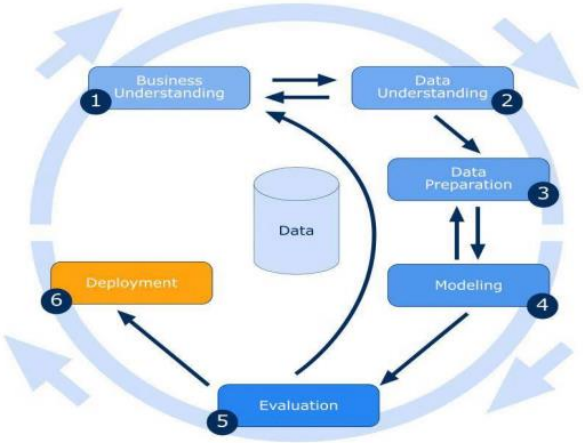
Dengan data yang terdistribusi, model mining juga harus terdistribusi agar efisien [6] memperkenalkan metode "*Distributed Bayesian Matrix Decomposition*" yang mampu mengatasi

noise heterogen serta mengoptimalkan pemrosesan data yang tersebar di beberapa node. Pendekatan ini penting ketika dataset terlalu besar untuk diolah pada satu mesin.

Dengan demikian, dalam penelitian ini, *Data Mining* berperan penting untuk menggali pola dari data teks mentah yang diperoleh dari media sosial X.. meunurut [7] proses kerangka kerja CRISP-DM *(Cross-Industry Standard Process for Data Mining)* adalah menyediakan pendekatan standar untuk proyek penambangan data. menemukan pola menarik atau informasi tersembunyi dari kumpulan data besar [8]. Teknik ini menggabungkan metode statistik, pembelajaran mesin, dan database. Tujuannya untuk mendukung pengambilan keputusan berbasis data. Dalam konteks media sosial, *Data Mining* digunakan untuk mengekstraksi wawasan dari teks tidak terstruktur.

2.2. Metodologi CRISP–DM

Kerangka kerja CRISP-DM yang telah banyak digunakan di berbagai studi untuk merancang alur analis data secara sistematis. Meskipunn dikembangkan sejak lama, CRISP-DM masih relavan hingga kini dan di adopsi secara luas dalam praktik *Data Mining* kontemporer [9], dengan mempertimbangkan struktur dan keunggulan metodologi CRISP-DM, kerangka ini digunakan dalam penelitian sebagai acuan dalam mengarahkan seluruh tahapan, mulai dari pemahaman data mentah di media sosial hingga tahap evaluasi hasil pengelompokan yang dibahas lebih lanjut pada bagian selanjutnya. CRISP-DM adalah kerangka kerja proses standar pada *Data Mining*. Dalam gambar 1 proses tersebut diilustrasikan menjadi enam tahap sebagai berikut:



Gambar 1. Metodologi CRISP – DM [10]

1. *Business Understanding*
   Tahapan ini bertujuan untuk memahami konteks bisnis atau sosial dari permasalahan yang akan diselesaikan melalui *Data Mining* [7]. Dalam penelitian ini, fokus utmaanya adalah memahami fenomena gagal bayar pinjaman onlne yang marak terjadi di masyarakat dan menimbulkan keresahan public. Permasalahan tersebut diangkat kembali menjadi dasar penelitian untuk merumuskan tujuan analisis data berbasis teks.
2. *Data Understanding*
   Tahap ini melibatkan pengumpulan dan eksplorasi data awal untuk pemahaman struktur, sumber, dan kualitas data. Data berupa *Tweet* yang dikumpulkan dari media sosial X (Twitter) Menggunakan keyword seperti “Pinjol”, “Galbay”, “Tagihan”, dan yang lainnya, setelah dikumpulkan data analis scara deskriptif untuk melihat distribusi dan bentuk teks.
3. *Data Preparation*
   Merupakan tahap transformasi data adar siap untuk dimodelkan. Proses ini meliputi pembersihan teks *(Cleaning), Case Folding, Tokenisasi, Stopword Removal, stemming,* dan

konversi ke dalam bentuk numerik menggunakan TF – IDF. Tahap ini sangat penting karena kualitas hasil model sangat tergantung pada kualitas data yang telah dipersiapkan.

4. *Modelling*
   Pada tahap ini dilakukan penerapan algoritma Machine Learning, penelitian ini menggunakan algoritma *K – Means Clustering* karakteristik data teks yang tidak memiliki label. Jumlah klister ditentukan menggunakan *Elbow Method* dan hasil pengelompokan di evaluasi menggunakan metrik evaluasi *Culustering.*
5. *Evaluation*
   Tahap ini bertujuan untuk mengevaluasi seberapa baik model yang dibangun bisa merepresentasikan struktur data. Dalam penelitian ini, evaluasi dilakukan melalui nilai Silhouette Score dan interpretasi hasil word cloud pada tiap cluster untuk melihat koherensi tema yang muncul.
6. *Deployment*
   Merupakan tahap penyajian hasil dalam bentuk visualisasi dan penarikan kesimpulan. Output yang dihasilkan berupa *word cloud*, daftar kata dominan tiap klaster, serta interpretasi yang mengarahkan pada identifikasi akar permasalahan gagal bayar pinjaman *Online*. Temuan penelitian ini berpotensi menjadi dasar bagi studi lanjutan maupun perumusan kebijakan berbasis data yang lebih adaptif terhadap dinamika sosial digital.

2.3. Representasi Teks dengan TF-IDF

TF-IDF adalah pendekatan statistik yang digunakan secara luas dalam sistem pemrosesan bahasa alami untuk merepresentasikan teks ke dalam bentuk numerik. Algoritma ini bekerja dengan cara mengukur seberapa penting suatu kata dalam dokumen relatif terhadap kumpulan dokumen lainnya [11] . TF-IDF terdiri dari dua komponen utama, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). Dalam konteks representasi teks, [12] membuktikan bahwa *Term Frequency–Inverse Document Frequency* (TF–IDF) adalah metode vektorisasi yang efisien dalam menonjolkan istilah penting dan mengurangi kata umum. Hal ini mendukung pemilihan TF–IDF sebagai dasar pengelompokan data pada penelitian ini

TF-IDF bekerja dengan menghitung bobot suatu kata berdasarkan seberapa sering kata tersebut muncul dalam satu dokumen dibandingkan dengan keseluruhan dokumen dalam korpus. Semakin sering suatu kata muncul dalam dokumen tersebut, maka nilai TF-nya semakin tinggi, yang mengindikasikan bahwa kata tersebut kemungkinan penting dalam konteks isi dokumen tersebut. Namun demikian, tidak semua kata yang sering muncul dalam satu dokumen selalu penting secara keseluruhan. Jika suatu kata muncul di banyak dokumen, maka nilai IDF-nya menjadi rendah, yang berarti kata tersebut kurang spesifik atau tidak terlalu membedakan antar dokumen. Sebaliknya, jika suatu kata jarang muncul di dokumen lain, maka IDF-nya tinggi, menunjukkan bahwa kata tersebut membawa informasi yang lebih khas atau unik untuk dokumen tempat ia muncul [11] .

2.4. *K-Means Clustering*

Algoritma *unsupervised learning* berbasis *centroid* yang membagi data ke dalam sejumlah berdasarkan jarak *Euclidean* terdekat antara titik data dan pusat klaster [13], [14]. Prosesnya dimulai dari inisialisasi centroid secara acak, kemudian dilakukan iterasi pengelompokan dan pembaruan posisi centroid hingga mencapai konvergensi. Metode ini efisien untuk data besar dan klaster memungkinkan pengungkapan pola tersembunyi di antara klaster yang terbentuk.

Dalam [15] menunjukkan bahwa *K-Means Clustering* unggul untuk dataset berukuran besar dan dapat memberikan segmentasi yang terukur terhadap tema dalam teks. Penentuan jumlah klaster (nilai K) sangat memengaruhi hasil akhir. Oleh karena itu, metode *Elbow* digunakan untuk menentukan nilai K optimal berdasarkan penurunan *Within-Cluster Sum Of Squares* (WCSS), sementara *Silhouette Score* digunakan untuk mengevaluasi kualitas

pengelompokan [16], [17]. Dalam konteks penelitian ini, *K-Means* sesuai karena dapat menangani data teks dari media sosial Twitter yang tidak berlabel.

2.5. *Machine Learning*

Pendekatan dari cabang kecerdasan buatan (*Machine Learning)* telah mendefinisikan yang memungkinakan sistem belajar dari data tanpa diprogram secara eksplisit, kemampuan prediktif dan deksriptif yang dimiliki oleh *Machine Learning* bahwa teknologi tersebut akan menjadi isu strategis yang signifikan pada masa mendatang [18]. kemampuan tersebut memungkinkan pola – pola tersembunyi dalam data secara otomatis melalui proses komputasi, sehingga Keputusan yang diambil dapat didasarkan pada informasi yang objektif dan terukur [19], [20]. *Machine Learning* dipandang sebagai salah satu pilar utama dalam pengembangan sistem berbasis kecerdasan buatan dan tranformasi digital lintas sektor dan Machine Learning juga berfokus pada pembuatan model prediktif berdasarkan analis pola dari historis [21]. Penelitian ini didasarkan pada perkembangan terkini dalam bidang *Machine Learning* dan *Natural Language Processing* (NLP) yang telah banyak dimanfaatkan untuk analisis teks di media sosial. Menurut [18], *Machine Learning* memiliki kemampuan adaptif untuk mengenali pola tersembunyi dari data tanpa perlu pemrograman eksplisit, sehingga relevan digunakan dalam eksplorasi motif sosial berbasis teks.

2.6. *Natural Language Processing* (NLP)

NLP merupakan cabang AI yang memungkinkan komputer dapat memahami dan mengolah bahasa alami manusia. NLP sangat penting untuk mengolah teks tidak terstruktur di media sosial seperti X (Twitter) yang mengandung slang, singkatan, dan gaya bahasa informal [22]. NLP memungkinkan peneliti mengekstraksi opini publik, emosi, dan pola bahasa dari unggahan daring [23], sehingga dapat digunakan untuk mengidentifikasi faktor di balik gagal bayar pinjaman *online*.

Tahapan pre-processing seperti tokenisasi, *stopword removal*, *stemming*, dan *lemmatization* penting untuk normalisasi teks dan mengurangi kompleksitas data [24]. Setelahnya, teks diubah ke bentuk numerik menggunakan teknik vektorisasi seperti TF-IDF yang terbukti efektif untuk mengubah data teks menjadi fitur numerik yang dapat diproses oleh algoritma seperti *K-Means* .

2.7. *Text Preprocessing*

Text Pre Processing merupakan tahapan dasar dalam NLP yang bertujuan untuk membersihkan, menyederhanakan, dan menormalkan data teks sebelum digunakan pada proses analisis atau pemodelan [25]. *Pre Processing* teks sangat penting untuk mengubah data mentah yang tidak terstruktur menjadi bentuk yang dapat dimengerti oleh algoritma komputasional. Menurut yang menyatakan bahwa kualitas hasil klasifikasi dan *Clustering* sangat bergantung pada kualitas *Pre Processing* teks yang dilakukan [26]

Dalam penelitian ini, *Pre Processing* dilakukan terhadap data berupa Tweet dengan pendekatan bertahap. Setiap tahapan dirancang untuk menghilangkan unsur-unsur yang tidak relevan dan menyempurnakan struktur teks agar dapat diolah menjadi representasi numerik seperti TF-IDF.

2.8. Media Sosial sebagai Sumber Data Sosial

Data Twitter (X) adalah *Platform* mikroblog yang sering digunakan dalam penelitian sosial karena karakteristiknya yang terbuka, singkat, dan responsif. Tweet merupakan data tidak terstruktur yang membutuhkan NLP dan *Pre Processing* untuk dianalisis. Analisis Twitter berguna untuk mengetahui tren dan persepsi publik terhadap isu tertentu, Termasuk pinjaman *online*.

X berdasarkan laporan digital 2025 yang dirilis oleh We are social dan Meltwater, pada januari 2025, jumlah pengguna X di Indonesia mencapai sekitar 25,2 juta pengguna, angka ini menempatkan sebagai negara dengan jumlah pengguna x terbesar keempat didunia, setelah Amerika Serikat, Jepang, dan India. Setara dengan 8,8% dari total populasi di Indonesia data ini bisa menjadikan sample yang mengandung informasi penting tentang berbagai permasalahan nya, Penelitian ini dilakukan untuk mengidentifikasi motif – motif gagal bayar pengguna layanan pinjaman *Online* di Platform X dengan memanfaatkan metode metode pendekatan *text mining* berbasis NLP dan *Clustering.*

2.9. Pinjaman *Online* dan Permasalahan Gagal Bayar

Pinjaman *online* yang dikenal sebagai *Financial technology* (Fintech) *lending*, merupakan bentuk layanan keuangan berbasis digital yang memungkinkan pengguna memperoleh akses terhadap kredit tanpa tatap muka maupun dokumen fisik. Menurut [27], kemudahan dalam mengakses layanan, proses pencairan dana yang cepat, serta persyaratan administratif yang minim telah mendorong pertumbuhan eksponensial layanan pinjaman *online* di Indonesia dalam lima tahun terakhir.

Sementara itu, [28] serta [29] memberikan landasan empiris bahwa fenomena gagal bayar tidak hanya bersumber dari faktor ekonomi, tetapi juga dipengaruhi oleh perilaku sosial dan tekanan psikologis pengguna. Justifikasi literatur ini memperkuat arah penelitian yang berupaya mengaitkan hasil clustering dengan dimensi perilaku dan motif sosial, bukan semata analisis numerik risiko kredit. Menurut penelitian dari [3] mengungkap bahwa tekanan ekonomi, seperti naiknya harga kebutuhan pokok dan rendahnya pendapatan, menjadi faktor dominan penyebab masyarakat mengandalkan pinjaman *online*. Fenomena ini turut diperparah oleh praktik gali lubang tutup lubang, di mana pengguna berutang dari satu Platform untuk membayar utang di *Platform* lain, sehingga memperbesar risiko gagal bayar.

Pertumbuhan pinjaman *online* juga dibarengi dengan meningkatnya permasalahan gagal bayar. Menurut studi oleh [1] salah satu penyebab utama gagal bayar adalah rendahnya literasi keuangan dan impulsivitas konsumen dalam mengambil keputusan keuangan secara cepat. Sementara itu, penelitian oleh[1], [30] menunjukkan bahwa tekanan ekonomi dan biaya hidup menjadi faktor signifikan yang mendorong masyarakat untuk meminjam, bahkan melebihi kapasitasnya.

Dalam [27], [31] menyoroti bahwa sebagian pengguna yang gagal bayar juga mengalami intimidasi atau penagihan tidak etis dari pihak penyedia pinjaman. Dengan mempertimbangkan kompleksitas masalah gagal bayar pada pinjaman *online*, pendekatan berbasis analisis teks dari media sosial menjadi penting untuk mengidentifikasi pola permasalahan tersebut secara otomatis. data Tweet yang dianalisis dalam penelitian ini difokuskan pada konten yang berkaitan dengan perilaku gagal bayar dan penyebab sosial-ekonominya.

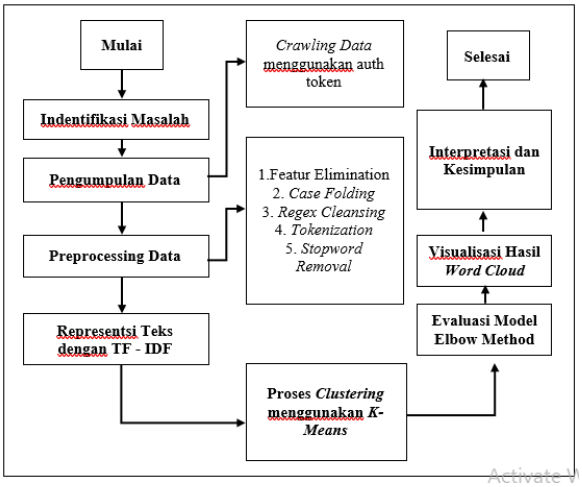
3. METODE PENELITIAN

Dalam penelitian ini dilakukan tahapan awal higga akhir penelitian yang dirancang sistematis mulai dari identifikasi permasalahan hingga interpretasi dan penarikan kesimpulan seperti yang ditunjukkan pada gambar 2. Setiap tahapan disusun berdasarkan pendekatan analisis data eksploratif berbasis teks yang bersumber dari media sosial, dengan menerapkan metode pengelompokan data menggunakan algoritma *K-Means Clustering* dan representasi fitur teks berbasis TF-IDF.

3.1. Identifikasi Masalah

Penelitian ini diawali dengan proses identifikasi masalah berdasarkan fenomena sosial digital terkait meningkatnya kasus gagal bayar pinjaman *online* yang muncul dalam percakapan publik di media sosial [1]. Permasalahan ini diamati dari tren konten pengguna yang mengeluhkan

penagihan tidak etis, penyalahgunaan data, atau penipuan berbasis layanan keuangan digital [3], [27] Proses ini penting untuk merumuskan fokus penelitian dan tujuan yang ingin dicapai, yakni mengelompokkan motif gagal bayar secara eksploratif menggunakan pendekatan *Machine Learning*.



Gambar 2. Tahapan Penelitian

3.2. Pengumpulan Data

Data dikumpulkan dari platform media sosial Twitter (X) dengan menggunakan proses *crawling* berbasis *auth token* [32]. Proses ini memungkinkan pengambilan data teks (tweet) yang relevan berdasarkan kata kunci seperti “pinjol”, “galbay”, dan “gagal bayar”. Data dikumpulkan dalam format CSV untuk kemudian diproses secara lebih lanjut. Proses pengumpulan dilakukan secara bertahap untuk menghindari pembatasan server Twitter yang berlaku secara periodik.

3.3.Pre Processing Data

Tahapan ini melibatkan pembersihan dan normalisasi data teks agar dapat diolah secara numerik. Proses yang dilakukan meliputi feature elimination (penghapusan kolom tidak relevan), *Case Folding* (konversi huruf menjadi huruf kecil), *regex cleansing* (penghapusan simbol dan karakter khusus), *tokenization* (pemecahan kalimat menjadi kata), serta *stopword removal* (penghapusan kata tidak bermakna). Tahapan ini bertujuan untuk meningkatkan kualitas input sebelum vektorisasi [33], [34].

3.4. Representasi Teks dengan TF-IDF

Teks yang telah diproses diubah menjadi bentuk numerik menggunakan teknik TF-IDF. Dalam [33] TF-IDF berfungsi memberikan bobot terhadap kata-kata dalam tweet berdasarkan frekuensi lokal dan globalnya. Representasi ini memungkinkan model *Machine Learning* mengenali perbedaan signifikan antar dokumen berdasarkan istilah pentingnya.

3.5. Proses *Clustering* menggunakan *K-Means*

Data vektor hasil TF-IDF digunakan sebagai input untuk algoritma *K-Means Clustering* [35]. Algoritma ini akan mengelompokkan tweet ke dalam sejumlah klaster berdasarkan kemiripan kata dan jarak antar dokumen dalam ruang vektor. Jumlah klaster (K) ditentukan melalui proses evaluasi model.

3.6. Evaluasi Model

Model K-Means dievaluasi dengan menggunakan dua pendekatan utama, yaitu Elbow Method. Dari penelitian [16], [36] *Elbow Method* digunakan untuk menentukan nilai K optimal dengan mengamati titik tekuk pada grafik inertia, sedangkan *Silhouette Score* digunakan untuk menilai seberapa baik setiap data sesuai dengan klaster yang ditempatinya.

3.7. Visualisasi Hasil (*Word Cloud*)

Setelah klaster terbentuk, dilakukan visualisasi hasil dengan menggunakan *Word Cloud* yang menampilkan kata-kata dominan pada setiap klaster. Visualisasi ini membantu interpretasi tema atau motif yang muncul dalam kelompok tweet.

3.8. Interpretasi dan Kesimpulan

Tahap akhir berupa interpretasi hasil *Clustering* yang diperoleh dan penarikan kesimpulan mengenai motif-motif gagal bayar pinjaman *online* berdasarkan data sosial digital [37]. Kesimpulan yang dihasilkan bersifat eksploratif dan menjadi kontribusi ilmiah terhadap kajian sosial digital dan penerapan Data Mining berbasis teks.

4. HASIL DAN PEMBAHASAN

Deskripsi data pada penelitian ini diperoleh melalui proses scraping media sosial Twitter (X) pada tahun 2023 menggunakan metode crawling berbasis auth token yang dikombinasikan dengan tools Python dan pustaka Tweet-Harvest. Proses pengambilan data dilakukan berdasarkan beberapa kata kunci seperti "pinjol", "gagal bayar", "galbay", "alasan pinjol", dan "korban pinjol". Hasil crawling disimpan dalam beberapa file CSV terpisah, seperti alasanpinjol.csv, galbay_pinjol.csv, dan lainnya, yang kemudian digabungkan menjadi satu dataset utama.

Dataset hasil penggabungan terdiri dari sekitar 2607 tweet. Atribut awal yang tersedia meliputi kolom ‘conversation_id_str’,‘created_at’,‘favorite_count’, ‘full_text’, ‘id_str’, ‘image_url’,‘in_reply_to_screen_name’,‘lang’,‘location’,‘quote_count’,‘reply_count’,‘retweet_count’,‘user_id_str’, ‘username’, dalam proses analisis, hanya kolom teks (full_text) yang digunakan karena memuat informasi utama dari opini pengguna. Sebelum diketahui distribusi jumlah tweet berdasarkan kata kunci yang digunakan, proses selanjutnya yang dilalukan adalah scraping data mentah secara langsung dari platform media sosial X (Twitter). Pada tabel 1 merupakan data yang diperoleh melalui proses *crawling* dengan menggunakan bahasa pemrograman Python dan pustaka Tweet-Harvest. Pengambilan data dilakukan secara autentikasi menggunakan token pengguna yang diperoleh melalui inspeksi elemen situs web, yang memungkinkan akses terhadap data publik tanpa menggunakan API resmi.

Tabel 1. Kata Kunci Crawling data

Kata Kunci	Jumlah Tweet hasil Scraping
Pinjol	404
Galbay	138
Paylater	402
Alasanpinjol	504
Galilubang	272
Korbanpinjol	210
Debtcollector	348
Terpaksapinjol	234
Kasusgalbay	95

Ketika autentikasi berhasil dikonfigurasi dengan menyisipkan token akses pengguna, Dalam penelitian ini, data publik diperoleh berdasarkan kata kunci yang telah ditentukan sebelumnya, dengan memanfaatkan pustaka Tweet-Harvest yang memungkinkan pengambilan

konten secara langsung dari laman Twitter tanpa menggunakan API resmi. Setiap pencarian diarahkan pada kata kunci tertentu dan dibatasi berdasarkan waktu, agar relevansi data tetap terjaga. sebelum dilakukan *feature elimination*, data memiliki sejumlah kolom seperti: [id, username, date, full_text, like_count, retweet_count]. Namun, setelah melalui tahapan *Pre Processing* secara menyeluruh—yang mencakup *Case Folding*, *tokenisasi*, penghapusan *stopword*, dan pembersihan teks menggunakan ekspresi reguler—data yang digunakan dalam proses analisis hanya menyisakan satu kolom utama yang telah diproses, yaitu hasil akhir dari variabel *stopword_removal*. Kolom ini berisi teks yang telah dibersihkan dan disederhanakan, dan menjadi representasi utama dari isi tweet yang dipertahankan untuk dianalisis lebih lanjut.

4.1. Hasil Vektorisasi TF-IDF

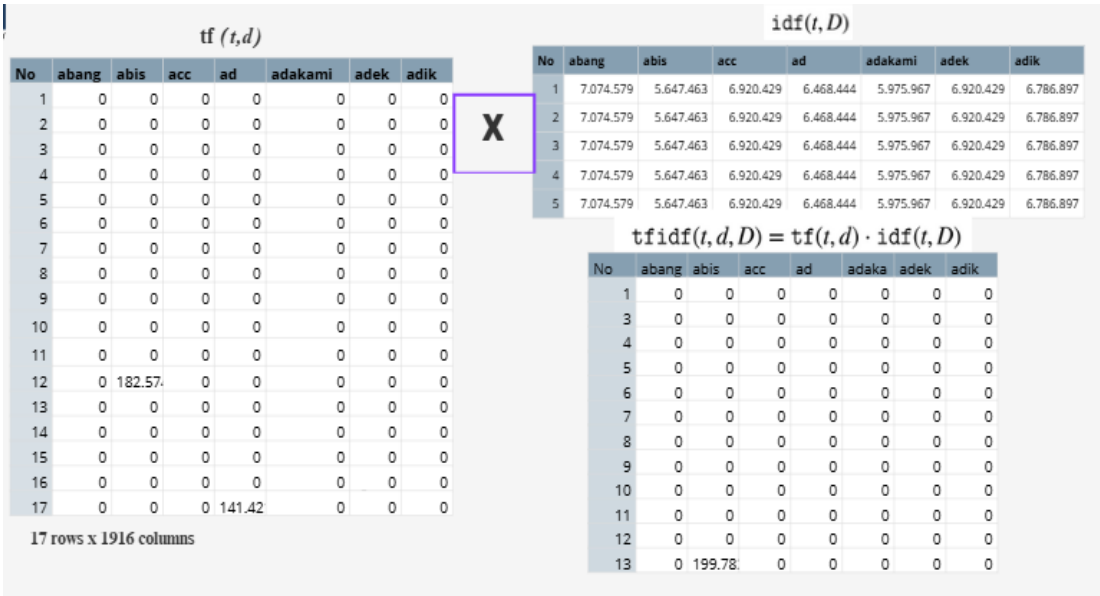
Vektorisasi teks dilakukan untuk mengubah data mentah berupa teks menjadi bentuk numerik, sehingga dapat dianalisis secara matematis dan digunakan sebagai input pada algoritma *Machine Learning*. Teknik yang digunakan pada penelitian ini adalah TF-IDF. Metode ini bekerja dengan menghitung bobot pentingnya suatu kata dalam satu dokumen dibandingkan dengan seluruh dokumen dalam korpus.

1. Matriks Term Frequency (TF)

TF merupakan representasi dari seberapa sering suatu kata (term) muncul dalam satu dokumen. Semakin sering suatu kata muncul dalam satu dokumen, maka nilai TF-nya akan semakin tinggi. Dalam penelitian ini, nilai TF dihitung dengan menggunakan pendekatan frekuensi absolut, yaitu jumlah kemunculan suatu kata dalam satu dokumen dibagi dengan jumlah seluruh kata dalam dokumen tersebut.

2. Matriks Inverse Document Frequency (IDF)

IDF digunakan untuk menyeimbangkan pengaruh kata-kata yang terlalu umum dan sering muncul di banyak dokumen. IDF akan memberikan bobot yang lebih tinggi terhadap kata-kata yang jarang muncul dalam keseluruhan dokumen, dan sebaliknya akan memberikan bobot lebih rendah pada kata-kata yang terlalu sering digunakan di banyak dokumen.



Gambar 3. Vektorisasi TF-IDF

3. Matriks TF – IDF

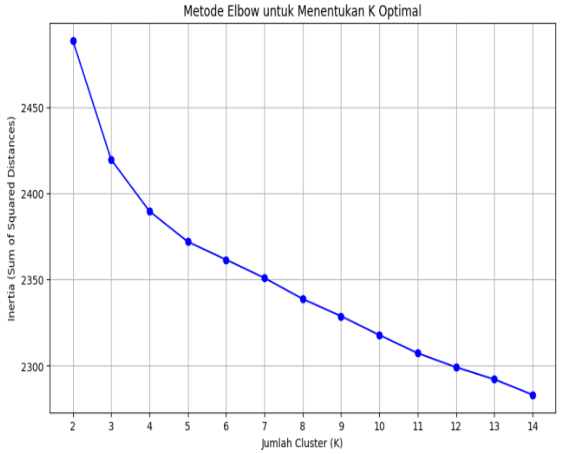
Matriks TF-IDF (Term Frequency – Inverse Document Frequency) merupakan hasil dari perkalian antara matriks TF dan matriks IDF. Nilai TF-IDF memberikan bobot yang seimbang antara frekuensi kemunculan kata dalam suatu dokumen dan tingkat keunikannya dalam korpus.

Semakin sering suatu kata muncul dalam satu dokumen, dan semakin jarang kata tersebut muncul di dokumen lainnya, maka nilai TF-IDF-nya akan semakin tinggi.

Data yang sudah melakukan pre processing akan lebih mudah dianalis pada tahapan vektorisasi TF-IDF yang ditunjukkan pada gambar 3, teks akan diubah kedalam bentuk numerik menggunakan TfidfVectorizer dari pustaka Scikit-Learn. Dengan parameter min_df = 5 dan max_df = 0.9, serta rentang ngram (1,2), diperoleh sebanyak 1916 fitur dari 2607 tweet yang tersedia. Fitur ini terdiri atas kata tunggal (unigram) dan kombinasi dua kata (bigram) yang memiliki frekuensi kemunculan yang signifikan di dalam korpus. Matriks TF-IDF yang dihasilkan memiliki bentuk (2607 × 1916), dan berbentuk sparse matrix karena sebagian besar nilai merupakan nol.

4.2. Penentuan Jumlah Cluster

Penentuan jumlah klaster optimal dalam proses pengelompokan dilakukan melalui pendekatan evaluatif menggunakan metode Elbow. Metode ini merupakan salah satu teknik yang banyak digunakan dalam analisis *Clustering*, dengan cara memantau perubahan nilai inertia (jumlah kuadrat jarak setiap titik terhadap pusat klasternya) terhadap variasi jumlah klaster (K). Inertia yang lebih rendah menunjukkan bahwa data lebih dekat dengan centroid masing-masing klaster, sehingga model lebih kompak dan representatif.



Gambar 4. Grafik Metode Elbow

Dalam penelitian ini, nilai K divariasikan dari 2 hingga 14. Untuk setiap nilai K dilakukan perhitungan inertia dan hasilnya dipetakan dalam bentuk grafik Elbow yang memperlihatkan hubungan antara jumlah klaster dan nilai inertia. Titik tekuk *(elbow point)* pada grafik menandai nilai K di mana penurunan inertia mulai melambat secara signifikan. Dengan kata lain, penambahan jumlah klaster setelah titik tersebut tidak lagi memberikan pengurangan inertia yang substansial. Berdasarkan gambar 4 yang diperoleh bahwa titik tekuk berada pada nilai k = 4. Oleh karena itu, jumlah klaster yang dipilih untuk proses *K-Means Clustering* ditetapkan sebanyak empat. Pemilihan ini didasarkan pada pertimbangan efisiensi model serta kemampuan interpretasi hasil yang optimal dalam konteks pengelompokan motif gagal bayar berdasarkan isi tweet.

4.3. Hasil Clustering

Berdasarkan hasil implementasi algoritma *K-Means* dengan parameter jumlah klaster (k) = 4 terhadap matriks TF-IDF yang telah dibentuk sebelumnya, diperoleh hasil pengelompokan sebanyak 2607 tweet ke dalam empat klaster yang berbeda. Proses pengelompokan dilakukan dengan pendekatan *Unsupervised Learning*, di mana setiap tweet diklasifikasikan ke dalam klaster berdasarkan kemiripan distribusi kata-katanya terhadap pusat klaster (*centroid*) yang ditentukan secara iteratif oleh model.

```
--- Hasil Klasterisasi (k=4) ---
Total tweet yang diproses: 2607
```

	cluster_label	Jumlah Tweet	Presentase (%)
0	0	306	11.74
1	1	439	16.84
2	2	1610	61.76
3	3	252	9.67

Gambar 5. Hasil Cluster

Distribusi jumlah tweet pada masing-masing klaster ditampilkan pada gambar 5, klaster 2 merupakan klaster yang memiliki jumlah anggota terbanyak, yaitu sebanyak 1610 tweet atau sekitar 61,76% dari total keseluruhan data. Hal ini menunjukkan bahwa sebagian besar tweet dalam dataset memiliki kemiripan pola kata yang terkonsentrasi pada tema dominan yang terklasifikasi dalam klaster tersebut. Klaster 1 berada di urutan kedua dengan 439 tweet (16,84%), diikuti oleh Klaster 0 dengan 306 tweet (11,74%), dan Klaster 3 dengan jumlah paling sedikit yaitu 252 tweet (9,67%).

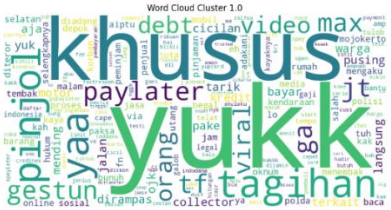
Distribusi yang tidak merata ini menggambarkan bahwa topik atau motif yang dibahas oleh pengguna dalam media sosial X (Twitter) cenderung terpusat pada satu tema utama, sedangkan tiga klaster lainnya mengelompokkan tema minor atau khusus. Analisis lebih lanjut terhadap kata-kata dominan dan contoh tweet dari masing-masing klaster akan disajikan pada subbab selanjutnya guna mengetahui makna semantik dari tiap kelompok hasil klasterisasi.

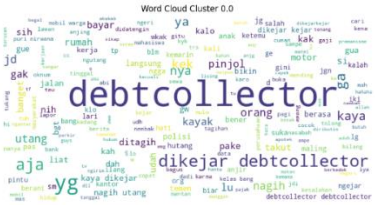
4.4. Interpretasi Hasil Clustering Berdasarkan Word Cloud

Berdasarkan hasil pengelompokan menggunakan algoritma *K-Means* dan evaluasi jumlah klaster optimal dengan metode Elbow, telah ditetapkan bahwa data tweet diklasifikasikan ke dalam empat klaster (k = 4). Dari hasil distribusi jumlah tweet per klaster, terlihat bahwa klaster nomor 2 merupakan klaster yang paling dominan, dengan proporsi sebesar 61,76% dari total keseluruhan data. Dominasi ini menunjukkan bahwa sebagian besar opini atau pengalaman pengguna berkonsentrasi pada tema atau isu yang terkandung dalam klaster tersebut.

Dalam proses interpretasi lebih lanjut, penulis menyusun pembahasan klaster secara berurutan mulai dari yang paling dominan (klaster 2), kemudian klaster 1, klaster 0, dan terakhir klaster 3. Penyusunan ini bertujuan untuk memberikan fokus terhadap kelompok motif gagal bayar yang paling banyak diungkapkan oleh pengguna, sekaligus memudahkan pembaca dalam memahami ragam isu yang berhasil diekstraksi dari data sosial digital.

Tabel 2. Hasil *Wordcloud cluster*

Cluster	Wordcloud
2	
1	

Cluster	Wordcloud
0	
3	

Setiap klaster dianalisis secara semantik menggunakan visualisasi *Word Cloud* yang ditujjukan pada tabel 2, dibentuk dari bobot TF-IDF pada kata-kata yang terdapat pada masing-masing kelompok. *Word Cloud* memungkinkan visualisasi representatif dari kata-kata yang paling informatif dalam suatu klaster, di mana ukuran huruf mencerminkan tingkat dominansi kata dalam teks.

Visualisasi klaster 0 menyoroti kata-kata yang dominan yaitu “debtcollector”, “ditagih”, “tagihan”, “hutang”, dan “dikejar” yang menunjukkan pengalaman pengguna yang menghadapi tekanan langsung maupun ancaman dari pihak penagih. Fenomena ini sesuai dengan temuan [28] yang menyebutkan bahwa praktik penagihan tidak etis sering melibatkan kekerasan verbal maupun penyebaran data pribadi, menciptakan tekanan psikologis terhadap debitur.

Klaster 1 menggambarkan narasi promosi, gestun (gesek tunai), dan ajakan viral yang berorientasi konsumtif. Hal ini dikarenakan kata dominan dari kluster 1 berupa “khusus, “yuk”, “paylater”, “viral”, “tagihan”, dan “pinjol”. Pola ini menunjukkan adanya praktik pemasaran agresif dan persepsi publik terhadap kredibilitas penyedia layanan. Temuan ini memperkuat analisis [29] bahwa rendahnya transparansi dan promosi menyesatkan pada *platform FinTech* dapat menurunkan kepercayaan masyarakat.

Visualisasi diatas juga memperlihatkan bahwa klaster 2 diwarnai oleh kata-kata seperti “pinjol”, “yukk”, “khusus”, “tagihan”, “gestun”, “collector”, “bayar”, “video”, “viral”, dan “cicilan”. Kata-kata promosi seperti “yukk”, “khusus”, dan “max”, serta istilah teknis seperti “gestun”, “paylater”, dan “credit” ditemukan secara mencolok, yang menunjukkan bahwa klaster ini telah dipenuhi oleh konten promosi terkait pinjaman *online*. Promosi tersebut umumnya disebarkan oleh akun-akun yang menawarkan jasa pencairan dana atau pelunasan utang secara cepat.

Klaster 3 menampilkan pola klasik “gali lubang tutup lubang” yang menunjukkan strategi bertahan jangka pendek akibat tekanan finansial. Pola ini sejalan dengan teori *Behavioral Economics* [38] yang menjelaskan bahwa individu dengan tekanan ekonomi tinggi cenderung mengambil keputusan keuangan irasional dan berisiko. Fenomena ini juga diperkuat oleh [28] yang menyoroti rendahnya literasi keuangan sebagai penyebab utama terjebaknya pengguna dalam siklus pinjaman berulang.

Selain itu, kemunculan kata “video”, “viral”, dan “OJK” mengindikasikan bahwa kampanye pemasaran atau testimoni telah diviralkan secara sengaja untuk menarik perhatian publik. Kata-kata seperti “collector”, “tagihan”, dan “dirampas” juga menunjukkan adanya pengalaman intimidasi yang dirasakan oleh pengguna setelah terlibat dalam layanan tersebut.

Keempat kluster diatas membentuk satu kesatuan pemahaman bahwa perilaku gagal bayar dalam pinjaman daring bukan hanya akibat faktor ekonomi, tetapi juga hasil interaksi sosial, budaya konsumtif, dan bias perilaku yang kompleks.Tabel 4 merupakan hasil analisis pengelompokan berkatitan dengan isu topik.

Tabel 3. Hasil Isu Topik

Topik	Motif Gagal Bayar
Topik 1	Pengaruh promosi gesek tunai dan Tekanan ekonomi
Topik 2	Manipulasi Sosial oleh Jasa Pelunasan Abal-abal
Topik 3	Dampak Emosional dan Ancaman DebtCollector
Topik 4	Gaya hidup, gengsi, gali lubang tutup lubang

5. KESIMPULAN

Penelitian ini dilakukan untuk mengidentifikasi motif-motif gagal bayar (galbay) pada pengguna pinjaman *online* dengan memanfaatkan data percakapan masyarakat di media sosial X (Twitter). Dengan menerapkan pendekatan *Unsupervised Learning* menggunakan representasi TF-IDF dan algoritma *K-Means Clustering*, diperoleh pengelompokan data teks ke dalam beberapa klaster berdasarkan kesamaan kata yang muncul dalam tweet.

Hasil klasterisasi menunjukkan bahwa motif gagal bayar yang paling dominan berkaitan erat dengan tekanan ekonomi serta pengaruh promosi yang menyesatkan, seperti penawaran gesek tunai (gestun) yang tersebar luas di media sosial. Klaster dominan memperlihatkan bahwa sebagian besar pengguna terdorong mengambil pinjaman bukan karena kebutuhan mendesak, melainkan karena godaan akses cepat dan kemudahan proses pencairan yang sering kali dipromosikan tanpa transparansi risiko.

Tekanan ekonomi yang meliputi kebutuhan hidup harian, minimnya penghasilan tetap, serta kecenderungan mencari pemasukan instan melalui praktik judi *online*, menjadi latar belakang utama dari tindakan gagal bayar. Promosi agresif seperti testimoni palsu, layanan "bantu pelunasan", hingga tawaran gesek tunai turut memperparah situasi, karena mendorong pengguna untuk mengambil keputusan finansial tanpa pertimbangan jangka panjang.

Temuan ini mengindikasikan bahwa gagal bayar pada pinjaman *online* bukan hanya permasalahan individu, tetapi juga hasil dari sistem promosi yang eksploitatif dan lemahnya ketahanan ekonomi masyarakat. Sementara itu, klaster lainnya menunjukkan adanya pengaruh dari promosi menyesatkan, tekanan penagihan yang bersifat agresif, dan gaya hidup konsumtif yang mendorong siklus utang yang terus berulang. Temuan ini membuktikan bahwa pendekatan pengelompokan berbasis teks dari media sosial dapat memberikan gambaran tematik yang lebih dalam terhadap perilaku dan motif pengguna pinjaman *online*. Metode TF-IDF dan K-Means terbukti efektif untuk menggali informasi tersembunyi dari data yang tidak berlabel, serta mampu mengekstrak pemetaan wacana digital terkait fenomena sosial yang kompleks seperti gagal bayar.

Dengan demikian, kontribusi utama penelitian ini terletak pada pergeseran fokus analisis risiko kredit dari data terstruktur yang konvensional menuju pemanfaatan data tidak terstruktur berbasis media sosial untuk memahami motif sosial dan perilaku gagal bayar. Pendekatan ini memperluas cakupan kajian risiko keuangan dengan mempertimbangkan dimensi perilaku dan sosial yang selama ini kurang tereksplorasi dalam analisis risiko kredit tradisional.

6. SARAN

Penelitian ini membuka ruang untuk eksplorasi lebih lanjut, khususnya dalam hal pendekatan metode dan perluasan cakupan data. Oleh karena itu, beberapa saran yang dapat disampaikan yaitu penelitian ini menggunakan K-Means sebagai metode pengelompokan karena kesederhanaan dan efisiensinya untuk data berdimensi tinggi seperti TF-IDF. Namun, untuk data sosial media yang bersifat sangat variatif dan tidak selalu membentuk distribusi klaster yang jelas, pendekatan lain seperti DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) atau LDA (*Latent Dirichlet Allocation*) dapat digunakan untuk menggali struktur tema yang lebih fleksibel atau berdasarkan probabilitas topik.

Dengan membandingkan performa hasil dari beberapa metode tersebut, keakuratan dan kedalaman interpretasi dapat ditingkatkan. Selain Twitter, opini pengguna juga banyak ditemukan di platform lain seperti Facebook, TikTok, dan forum seperti Kaskus atau Reddit. Penelitian lanjutan dapat melakukan perbandingan lintas platform untuk mengevaluasi perbedaan motif dan persepsi pengguna berdasarkan karakteristik demografis dan budaya komunikasi di masing-masing media.

Sebagai penguatan hasil klasterisasi, dapat ditambahkan analisis sentimen (positif, negatif, netral) atau deteksi emosi (takut, marah, sedih, frustrasi). Ini memungkinkan peneliti mengaitkan motif galbay tidak hanya dari kata kunci dominan, tetapi juga dari dimensi emosional yang menyertainya.

Untuk mendapatkan validitas eksternal, hasil dari media sosial dapat dikombinasikan dengan data resmi seperti laporan OJK, BI, atau survei perilaku keuangan. Hal ini akan memberikan konteks yang lebih komprehensif antara persepsi pengguna dan data faktual penyedia pinjaman *online*.

Hasil penelitian ini dapat dijadikan dasar untuk membangun sistem berbasis *Machine Learning* yang mampu memantau percakapan daring dan mendeteksi potensi gagal bayar secara dini. Hal ini dapat bermanfaat bagi lembaga keuangan dalam melakukan mitigasi risiko secara lebih proaktif dan beretika.

DAFTAR PUSTAKA

[1] W. H. Susanto and A. F. Chawa, "Aksi Gagal Bayar pada Perusahaan Fintech," *JSSH (Jurnal Sains Sosial dan Humaniora)*, vol. 5, no. 1, p. 9, Apr. 2021, doi: 10.30595/jssh.v5i1.9305.

[2] A. Werdayanti, "Jurnal Pendidikan Ekonomi," *Jurnal Pendidikan Ekonomi Vol 2 No.2 Juli, Tahun 2008*, vol. 3, no. 2, pp. 79–92, 2023.

[3] R. A. E. Wahyuni and B. E. Turisno, "Praktik Finansial Teknologi Ilegal Dalam Bentuk Pinjaman Online Ditinjau Dari Etika Bisnis," *Jurnal Pembangunan Hukum Indonesia*, vol. 1, no. 3, pp. 379–391, 2019, doi: 10.14710/jphi.v1i3.379-391.

[4] N. putu M. dewi P. Asti, "Upaya Hukum Otoritas Jasa Keuangan (OJK) dalam Mengatasi Layanan Pinjaman Online Ilegal," *Acta Comitas*, vol. 5, no. 1, p. 111, Apr. 2020, doi: 10.24843/AC.2020.v05.i01.p10.

[5] Z. Gustiana, "Exploration of Data Mining Techniques in Business Decision-Making," *Dharmawangsa: International Journal of the Social Sciences, Education and Humanitis*, vol. 5, no. 2, pp. 103–112, 2024, doi: 10.46576/ijsseh.v5i2.4647.

[6] C. Zhang, Y. Yang, W. Zhou, and S. Zhang, "Distributed Bayesian Matrix Decomposition for Big Data Mining and Clustering," *IEEE Trans Knowl Data Eng*, vol. 34, no. 8, pp. 3701–3713, 2022, doi: 10.1109/TKDE.2020.3029582.

[7] A. Pambudi, "Penerapan CRISP-DM Menggunakan MLR K-Fold Pada data Saham PT. Telkom Indonesia (Persero) TBK (TLKM) (Studi Kasus: Bursa Efek Indonesia Tahun 2015-2022)," *Jurnal Data Mining dan Sistem Informasi*, vol. 4, no. 1, p. 1, Mar. 2023, doi: 10.33365/jdmsi.v4i1.2462.

[8] C. S. Octiva, T. I. Fajri, E. B. Sulistiarini, S. Suharjo, and U. W. Nuryanto, "Penggunaan Teknik Data Mining untuk Analisis Perilaku Pengguna pada Media Sosial," *Jurnal Minfo Polgan*, vol. 13, no. 1, pp. 1074–1078, Jul. 2024, doi: 10.33395/jmp.v13i1.13936.

[9] A. Ardhi Baskara, N. Maharani Piranti, and M. Fahrury Romdendine, "Framework Data Mining: Sebuah Survei," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 3, pp. 4886–4895, May 2025, doi: 10.36040/jati.v9i3.13803.

[10] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," *Journal of Applied Informatics and Computing*, vol. 5, no. 2, pp. 103–108, 2021, doi: 10.30871/jaic.v5i2.3200.

[11] Maulidya Prastita Syah, Ajeng Puspa Wardani, Mohammad Idhom, and Trimono, "Perbandingan Representasi Teks Tf-Idf Dan Bert Terhadap Akurasi Cosine Similarity Dalam Penilaian Otomatis Jawaban Berbasis Teks," *Data Sciences Indonesia (DSI)*, vol. 5, no. 1, pp. 47–59, Jul. 2025, doi: 10.47709/dsi.v5i1.6021.

[12] H. Ma'rifah, A. Prasetya Wibawa, and M. I. Akbar, "Sains, Aplikasi, Komputasi dan Teknologi Informasi Klasifikasi artikel ilmiah dengan berbagai skenario preprocessing," vol. 2, no. 2, p. 70, 2020.

[13] N. Wakhidah, "Clustering Menggunakan K-Means Algorithm (K-Means Algorithm Clustering)," *Fakultas Teknologi informasi*, vol. 21, no. 1, pp. 70–80, 2014.

[14] R. Sibarani and O. Omby, "Algorithma K-Means Clustering Strategi Pemasaran Penerimaan Mahasisswa Baru Universitas Satya Negara Indonesia," *Jurnal Algoritma, Logika dan Komputasi*, vol. 1, no. 2, Nov. 2018, doi: 10.30813/j-alu.v1i2.1367.

[15] B. Hakim, F. Joanda Kaunang, C. Susanto, J. Salim, and R. Indradjaja, "Implementasi Machine Learning Dalam Pengelompokan Musik Menggunakan Algoritma K-Means Clustering," 2025. [Online]. Available: http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/indexBhustomyHakim|http://jom.fti.budiluhur.ac.id/index.php/IDEALIS/index|

[16] I. Herdiana, M. A. Kamal, Triyani, M. N. Estri, and Renny, "A More Precise Elbow Method for Optimum K-means Clustering," pp. 1–22, 2025.

[17] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index Pada Hasil Cluster K-Means Dan DBSCAN," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3S1.5001.

[18] P. Santoso, H. Abijono, and N. L. Anggreini, "Algoritma Supervised Learning dan Unsupervised Learning dalam Pengolahan DatA," *Unira Malang* |, vol. 4, no. 2, 2021.

[19] N. Hafidhoh, A. P. Atmaja, G. N. Syaifuddiin, I. B. Sumafta, S. M. Pratama, and H. N. Khasanah, "Machine Learning untuk Prediksi Kegagalan Mesin dalam Predictive Maintenance System," *Jurnal Masyarakat Informatika*, vol. 15, no. 1, pp. 56–66, May 2024, doi: 10.14710/jmasif.15.1.63641.

[20] R. G. Wardhana, G. Wang, and F. Sibuea, "Penerapan Machine Learning dalam Prediksi Tingkat Kasus Penyakit di Indonesia," 2023.

[21] I. Fitriyaningsih, Y. Basani, and L. M. Ginting, "Machine Learning: Prosperity of Rainfall, Water Discharge, and Flood With Web Application in Deli Serdang," *Jurnal Penelitian Komunikasi dan Opini Publik*, vol. 22, no. 2, Dec. 2018, doi: 10.33299/jpkop.22.2.1752.

[22] S. A. Azzahra, N. Wachid, and A. Majid, "Klasifikasi dan Analisis Semantik Cyberbullying Sosial Media X : Integrasi Web Scraping dan Natural Language Processing (NLP)," vol. 11, no. 2, pp. 353–360, 2025.

[23] Badry Ali Mustofa and Wawan Laksito Yuly Saptomo, "Use of Natural Language Processing in Social Media Text Analysis," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 4, no. 2, pp. 1235–1238, Feb. 2025, doi: 10.59934/jaiea.v4i2.875.

[24] D. Choirinnisa *et al*., "LDA Topic Modeling: Twitter-Based Public Opinion on Indonesian Ministry of Finance," *Sinkron*, vol. 9, no. 2, pp. 849–863, May 2025, doi: 10.33395/sinkron.v9i2.14719.

[25] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi," *IOP Conf Ser Mater Sci Eng*, vol. 874, no. 1, p. 012017, Jun. 2020, doi: 10.1088/1757-899X/874/1/012017.

[26] 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)," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 406, Apr. 2021, doi: 10.30865/mib.v5i2.2835.

[27] M. Yusuf Ibrahim, "Kewenangan Otoritas Jasa Keuangan untuk Mengatasi Layanan Pinjaman Online Illegal," *FENOMENA*, vol. 21, no. 2, p. 173, Nov. 2023, doi: 10.36841/fenomena.v21i2.3777.
[28] R. Ani, E. Wahyuni, and E. Turisno, "Praktik Finansial Teknologi Ilegal dalam Bentuk Pinjaman Online Ditinjau Dari Etika Bisnis."
[29] "3777-97-20032-1-10-20240822".
[30] J. Z. Y. Arvante, "Dampak Permasalahan Pinjaman Online dan Perlindungan Hukum Bagi Konsumen Pinjaman Online," *Ikatan Penulis Mahasiswa Hukum Indonesia Law Journal*, vol. 2, no. 1, pp. 73–87, 2022, doi: 10.15294/ipmhi.v2i1.53736.
[31] N. I. Rahmahafida, "Perlindungan Hukum Pihak Pemberi Pinjaman pada Layanan Pinjaman Pendidikan Berbasis Teknologi Informasi terhadap Risiko Gagal Bayar," *Jurist-Diction*, vol. 3, no. 2, p. 541, 2020, doi: 10.20473/jd.v3i2.18203.
[32] M. A. Syahira and R. Kurniawan, "Analisis Sentimen Cyberbullying Pada Media Sosial X Menggunakan Metode Support Vector Machine," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 3, p. 1724, Jul. 2024, doi: 10.30865/mib.v8i3.7926.
[33] H. Ma'rifah, A. P. Wibawa, and M. I. Akbar, "Klasifikasi Artikel Ilmiah Dengan Berbagai Skenario Preprocessing," *Sains, Aplikasi, Komputasi dan Teknologi Informasi*, vol. 2, no. 2, p. 70, Apr. 2020, doi: 10.30872/jsakti.v2i2.2681.
[34] R. Rinandyaswara, Y. A. Sari, and M. T. Furqon, "Pembentukan Daftar Stopword Menggunakan Term Based Random Sampling Pada Analisis Sentimen Dengan Metode Naïve Bayes (Studi Kasus: Kuliah Daring Di Masa Pandemi)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 4, p. 717, Aug. 2022, doi: 10.25126/jtiik.2022934707.
[35] R. S. Nurhalizah, R. Ardianto, and P. Purwono, "Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review," *Jurnal Ilmu Komputer dan Informatika*, vol. 4, no. 1, pp. 61–72, 2024, doi: 10.54082/jiki.168.
[36] B. Hakim, F. J. Kaunang, C. Susanto, J. Salim, and R. Indradjaja, "Implementasi Machine Learning dalam Pengelompokan Musik Menggunakan Algoritma K-Means Clustering," *IDEALIS : InDonEsiA journaL Information System*, vol. 8, no. 1, pp. 74–83, Jan. 2025, doi: 10.36080/idealis.v8i1.3357.
[37] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python," *Bina Insani Ict Journal*, vol. 7, no. 2, p. 156, 2020, doi: 10.51211/biict.v7i2.1422.
[38] "Prospect Theory: Decision Research 'A Branch Of Percep Iron Ics Dar ~ eI K:hn:man ii."