BAB II

LANDASAN TEORI

2.1 Teori tentang Topik Skripsi

2.1.1 *Financial Technology*

Teknologi digital yang sangat canggih telah berkembang ke arah yang lebih modern dan efisien, terutama di sektor keuangan [10]. Dalam sektor keuangan, memberikan perkembangan dalam teknologi sangat penting. Teknologi dan keuangan memiliki keterkaitan, Saat ini, teknologi finansial merupakan salah satu bentuk teknologi yang berkontribusi pada inovasi keuangan..

Menurut [11] *Financial Technology* adalah penerapan teknologi dalam sistem keuangan yang dapat menyediakan barang, jasa, teknologi, serta inovasi bisnis yang akan memengaruhi keseimbangan terkait nilai mata uang, keseimbangan terkait ekonomi, dan efektivitas keamanan dalam melakukan kerangka kerja cicilan. Proses transaksi keuangan tersebut dilakukan secara *peer to peer* untuk melakukan peminjaman dana secara digital agar lebih mudah bagi individu untuk mendapatkan kredit tanpa dibatasi oleh keberadaan.

2.1.2 Media Sosial

Media sosial adalah jenis media *online* yang memudahkan pengguna untuk berpartisipasi dan berbagi, memungkinkan interaksi sosial dalam jarak jauh [12]. Media sosial memberikan kemudahan yang diberikan bagi penggunanya untuk bertemu dengan orang yang tidak kenal sebelumnya. Dengan kata lain, media sosial dapat secara terbuka menyambut orang-orang untuk terlibat dengan membagikan pendapat mereka, baik melalui komentar atau menyebarkan informasi dengan cepat.

Selain itu, menggunakan media sosial untuk komunikasi *online* memiliki keuntungan yang meliputi [13]:

a. Membina hubungan baru sebagai hasil dari kemudahan orang asing dari seluruh dunia untuk berteman melalui media sosial.
b. Memotivasi diri sendiri untuk selalu mengembangkan keterampilan mereka melalui media sosial.
c. Media sosial dapat membuat kedekatan antar sesama untuk menjadi lebih bersahabat, empati dan juga perhatian.

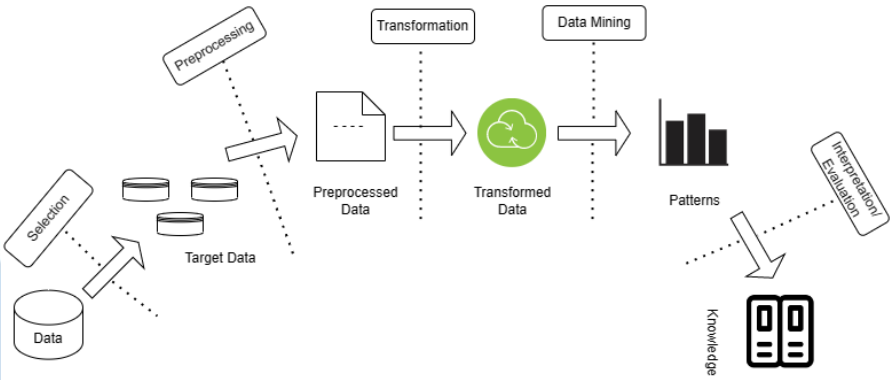
Dengan demikian, media sosial dapat menjadi sebuah *platform* yang menjembatani para penggunanya untuk lebih mudah dalam berinteraksi dan mengekspresikan diri mereka untuk mengemukakan pendapat atau memberikan sebuah *insight* baru agar didengar oleh orang lain.

2.2 Teori tentang Framework / Algoritma yang digunakan

2.2.1 *Text Mining*

Text mining adalah teknik mengidentifikasi pola dalam teks dan mengekstraknya sebagai laporan atau data yang dapat digunakan untuk mengubah pengaturan *ad hoc* menjadi segudang pola yang teratur [14]. Tujuan dari *text mining* adalah untuk mengidentifikasi kata-kata yang dapat menggambarkan konten suatu informasi, sehingga memungkinkan dilakukannya analisis terhadap keterhubungan informasi tersebut. Proses *text mining* biasanya secara luas dicirikan sebagai siklus informasi yang serius di mana *user* berinteraksi dengan bermacam-macam dokumen setelah beberapa waktu menggunakan alat analisis.

Proses kerja *text mining* sebagian besar dipengaruhi oleh penelitian *data mining* [15]. Namun, *text mining* berbeda dengan *data mining* karena *text mining* mengambil polanya dari basis data terstruktur, sedangkan *data mining* menggunakan struktur sumber data teks yang tidak terstruktur.



Gambar 2. 1 Tahapan *Text Mining* [16]

Pada Gambar 2.1 menunjukkan fase-fase siklus dalam *text mining* yang terdiri dari empat fase siklus, yaitu *preprocessed data*, *transformed data*, *Data Mining*, dan *evaluation* [16].

Fase pertama dalam siklus *text mining* adalah *preprocessed* data, atau dalam bahasa Indonesia disebut sebagai pra-pemrosesan data. *Preprocessing* data adalah proses persiapan dan pembersihan data teks sebelum dilakukan analisis atau pemodelan lebih lanjut [16]. Tujuan dari preprocessing data adalah untuk mengubah data teks mentah menjadi format yang lebih terstruktur dan mudah untuk diolah [16]. Tahap ini biasanya disebut dengan *case folding, tokenization*, dan juga pembersihan pada data untuk menghilangkan karakter khusus atau tanda baca.

Fase kedua dari siklus *text mining* adalah tahap *transformed data*. Pada tahap di mana data teks yang telah melalui tahap preprocessing diubah menjadi representasi numerik yang dapat diproses lebih lanjut [17]. Tujuan dari tahap ini adalah untuk mengubah data teks menjadi bentuk yang dapat dimengerti dan diproses oleh model atau algoritma pembelajaran mesin. Tahap ini biasanya meliputi perhitungan *Term Frequency-Inverse Document Frequency* (TF-IDF) [18].

Tahapan proses ketiga dalam *text mining* adalah *Data Mining*. *Data Mining* adalah proses ekstraksi informasi yang berharga, pola, dan pengetahuan dari data teks yang besar dan kompleks [19]. Tujuan

dari *Data Mining* dalam konteks *text mining* adalah untuk mengidentifikasi pola, tren, hubungan, dan wawasan yang tersembunyi dalam data teks [19]. Tahap ini biasanya meliputi pembentukan model atau pola yang dipakai dalam pemrosesan data.

Tahapan yang terakhir dalam *text mining* adalah *evaluation*. Tahapan *evaluation* digunakan untuk mengukur kinerja model atau algoritma yang telah dibangun dalam proses *data mining*. Melalui penggunaan metrik evaluasi yang sesuai, seperti akurasi, presisi, *recall*, *f1-score*, atau metrik khusus lainnya. Dalam proses ini dapat memahami sejauh mana model dapat mengklasifikasikan, mengelompokkan, atau menghasilkan prediksi yang akurat berdasarkan data teks [16].

Dalam berbagai proses yang digunakan dalam *text mining* tidak ada aturan yang baku mengenai proses seperti apa atau urutannya seperti apa yang digunakan dalam *text mining* tersebut, karena semua tergantung dari hasil *output* yang diinginkan dari data yang diolah. Berikut ini adalah penjelasan mengenai tahapan pemrosesan pada *text mining*:

A. *Case Folding*

Case folding adalah prosedur yang digunakan oleh *text pre-processing*. Tahap ini melibatkan standarisasi penggunaan huruf kapital, yang dapat menjadi masalah ketika mengkategorikan data dalam jumlah besar [20], sebagai contoh, data teks yang mungkin ditulis sebagai "PiNJAman ONLinE" akan ditransformasikan dengan menggunakan *case folding* menjadi huruf kecil (*lower case*) semua. Hal ini akan memastikan bahwa teks yang tidak konsisten akan digeneralisasi menggunakan huruf yang sama untuk semua karakter.

B. *Remove Duplicate*

Pada fase ini, sebuah siklus akan dijalankan untuk menghapus beberapa informasi yang memiliki kemiripan dengan informasi lain, tujuannya bahwa informasi dengan nilai yang sama tidak boleh terlalu sering diulang-ulang agar siklus untuk menangani sebuah informasi dapat berjalan dengan lancar.

C. *Filtering*

Langkah selanjutnya dalam proses *text mining* adalah penyaringan kata. Pada tahap ini, kata-kata yang memiliki tingkat signifikansi rendah dalam pemrosesan data akan dieliminasi. Selain itu, pada siklus ini akan dilakukan pembersihan pada beberapa informasi untuk menghapus simbol-simbol, *emoticon*, angka, dan yang lainnya untuk mengurangi *noise* [20], sehingga data tersebut tidak menyebabkan kesalahan dalam pengambilan keputusan.

D. *Tokenizing*

Pada tahap *Tokenizing* atau biasa yang disebut dengan tokenisasi berfungsi untuk proses pemecahan teks kalimat yang berbentuk panjang menjadi kata-kata yang disebut dengan *token* [20]. Selain itu, dengan tokenisasi dapat membedakan antara pemisah kata atau bukan, tujuannya adalah untuk memecah deskripsi, yang pada awalnya disusun sebagai kalimat akan diubah menjadi kata-kata dan meniadakan pembatas informasi seperti titik (.), spasi, dan angka.

Tabel 2. 1 Contoh *Tokenizing*

Contoh Tokenizing
['Ini', 'adalah', 'contoh', 'proses', 'memecah', 'kalimat', 'Tokenizing', 'adalah', kalimat', 'menjadi', 'token-token']

E. *Stopword*

Pada tahap *stopword* biasanya digunakan untuk membantu mengurangi jumlah kata dalam teks. Dalam *dataset* teks yang besar, kata-kata *stopwords* mungkin muncul sangat sering tetapi tidak memberikan banyak informasi tentang konten atau tujuan analisis. Dengan menghapus stopwords dapat mengurangi dimensi data dan fokus pada kata-kata yang lebih penting dan bermakna. Berikut merupakan contohnya:

Tabel 2. 2 Contoh *Stopwords*

Sebelum	Sesudah
Ini adalah contoh teks yang berisi beberapa kata yang umum dan tidak memberikan banyak informasi.	Ini contoh teks berisi kata-kata umum memberikan informasi.

Dalam contoh pada Tabel 2.2, *stopwords* seperti "adalah", "yang", "dan", "banyak", "ini", "beberapa", dan "tidak" dihapus dari teks. Hal ini dilakukan untuk menghilangkan kata-kata yang umum dan tidak signifikan sehingga fokus dapat ditujukan pada kata-kata yang lebih penting dalam pemrosesan teks.

F. *Lemmatization*

Pada tahap *lemmatization* biasanya digunakan untuk mengurangi jumlah indeks yang berbeda dalam satu set data sehingga kata dengan awalan atau akhiran dapat dikembalikan ke bentuk aslinya [21]. Tujuan utama lemmatisasi adalah untuk mereduksi kata-kata ke bentuk dasar yang memiliki makna yang sama. Ini membantu dalam normalisasi teks, di mana variasi kata yang berbeda (misalnya bentuk kata kerja dalam berbagai waktu, kata benda dalam bentuk jamak, dll.) dapat disederhanakan menjadi bentuk dasar yang konsisten.

Tabel 2. 3 Contoh *lemmatization*

Sebelum	Sesudah
['berlari', 'melompat', 'terbaik']	['lari', 'lompat', 'baik']

Dalam contoh pada Tabel 2.3, kata-kata asli dalam bahasa Indonesia telah diubah menjadi bentuk dasar mereka menggunakan lemmatisasi seperti berlari, melompat dan terbaik.

2.2.2 *Sentiment Analysis*

Metode yang dikenal sebagai analisis sentimen atau penggalian opini dilakukan untuk memastikan bagaimana sebuah perasaan disampaikan dalam sebuah teks. Analisis sentimen atau penggalian opini meliputi berbagai aspek seperti pemrosesan bahasa alami, linguistik komputasi, dan eksplorasi opini [21], tujuannya untuk menyelidiki pendapat, perasaan, evaluasi, sikap, penilaian, dan emosi seseorang.

Hal utama yang dilakukan dalam analisis sentimen adalah untuk mengkategorikan polaritas tekstual dalam data [21], sehingga pendapat dari masing-masing kategori yang memiliki nilai kategori kata positif, nilai kategori kata negatif, atau nilai kategori kata netral dapat diukur sejauh mana dampak dan kegunaan dari analisis sentimen yang dilakukan.

Dengan demikian, analisis sentimen dapat digunakan untuk memahami dan mengekstraksi sentimen atau opini yang terkandung dalam teks, baik itu dalam bentuk ulasan, *tweet*, komentar, atau yang lainnya.

2.2.3 *TF-IDF*

TF-IDF adalah singkatan dari "*Term Frequency-Inverse Document Frequency*" yang merupakan metode dalam pemrosesan data untuk menghitung bobot kata dalam suatu dokumen atau koleksi

dokumen. Teknik *TF-IDF* dapat digunakan buat menimbang hubungan antara kata atau istilah terhadap data [22]. Tahap ini akan mengukur sejauh mana kata-kata dalam dokumen memiliki bobot penting atau signifikansi relatif terhadap koleksi dokumen secara keseluruhan, seperti mengidentifikasi kata-kata yang paling penting dalam dokumen sampai menghitung bobot pada kata-kata dalam dokumen. Berikut ini adalah rumus untuk menghitung *Term Frequency*:

$$tf = 0{,}5 + 0{,}5 \text{ x } \frac{(tf)}{max\ (tf)} \quad \textbf{(2.1)}$$

Keterangan:

tf = Banyaknya kata yang dicari pada sebuah data.

max(tf) = Jumlah kemunculan terbanyak *term* pada data yang sama.

Selain itu, untuk menghitung *Inverse Document Frequency* (IDF) adalah sebagai berikut:

$$idf_t = \log\left(\frac{\mathrm{D}}{df_t}\right) \quad \textbf{(2.2)}$$

Keterangan:

Nilai D = Total data yang terkumpul

df_t = Jumlah data yang mengandung term *t*.

IDF = Inversed Document Frequency ($\log^2(\frac{(D)}{\mathrm{df}})$)

Oleh karena itu, metode ini bertujuan untuk menemukan representasi numerik dari setiap bagian data, yang menghasilkan pembuatan vektor antara data dan frekuensi yang ditunjukkan oleh representasi numerik dari istilah frekuensi dalam data. Jumlah bobot perhitungan yang dihasilkan akan meningkat, dan sebagai hasilnya, kemiripan data dengan frekuensi akan meningkat.

2.2.4 *Text Classification*

Text Classification adalah suatu teknik *machine learning* yang dapat digunakan untuk mengoordinasikan dan mengelompokkan hampir semua hal, termasuk dokumen, rekam medis, dan jenis teks lainnya [23]. Klasifikasi teks berfungsi untuk menawarkan kerangka kerja yang baik dalam pemrosesan data, karena biasanya digunakan untuk pengaplikasian untuk deteksi spam, kategorisasi berita, klasifikasi topik dan yang lainnya.

Cara kerja dari *text classification* dapat dilakukan dengan dua cara yaitu klasifikasi secara manual dan juga otomatis [24]. Pada klasifikasi manual digunakan untuk menafsirkan dan mengkategorikan konten teks pada *annotator* manusia, meskipun strategi ini biasanya dapat menghasilkan hasil yang berkualitas tinggi, namun strategi ini memakan waktu yang lama. Cara kerja dengan klasifikasi otomatis dapat menerapkan proses *machine learning*, *natural language processing* (NLP), dan teknik lainnya yang ditujukan untuk mengklasifikasikan teks dengan cara otomatis dan waktu yang diperlukan cukup singkat.

Secara umum, klasifikasi teks memuat empat tingkat ruang lingkup yang berbeda untuk diterapkan [25]:

1. Tingkat Dokumen

   Pada tingkat dokumen ini, metode akan memperoleh kategori yang relevan dari dokumen tersebut secara lengkap.

2. Tingkat Paragraf

   Pada tingkat paragraf ini, metode akan memperoleh kategori yang relevan berdasarkan bagian dari satu paragraf tersebut atau dengan kata lain setiap satu paragraf akan memperoleh sebagian dari isi dokumen tersebut.

3. Tingkat Kalimat

   Pada tingkat kalimat ini, diperoleh kategori yang relevan berdasarkan satu kalimat. Hal ini hampir mirip dengan tingkat paragraf, karena tingkat kalimat akan memperoleh sebagian dari paragraf tersebut.

4. Tingkat Sub-Kalimat

   Pada tingkat sub-kalimat dalam proses yang terakhir, metode akan memperoleh kategori satu atau beberapa ekspresi yang relevan untuk menentukan perhitungan model pada sebagian kalimat.

Dengan demikian, langkah terpenting dari klasifikasi teks adalah harus dapat memilih pengklasifikasi yang terbaik berdasarkan data yang dimiliki dan harus dapat memahami konseptual yang lengkap dari setiap metode yang dipakai, karena setiap metode semuanya bagus dan juga efektif. Klasifikasi teks merupakan metode *machine learning* yang paling umum digunakan bagi banyak penelitian, karena mudah dipahami [26].

2.2.5 *Support Vector Machine*

Support Vector Machine (SVM) adalah teknik pelatihan akademik lanjutan yang dimanfaatkan untuk melakukan analisis data dan mengidentifikasi pola atau struktur yang relevan dalam konteks klasifikasi dan analisis regresi [27]. *Support Vector Machine* memiliki gagasan yang lebih berkembang dan berbeda dalam pemodelan klasifikasi jika dibandingkan dengan teknik klasifikasi lainnya.

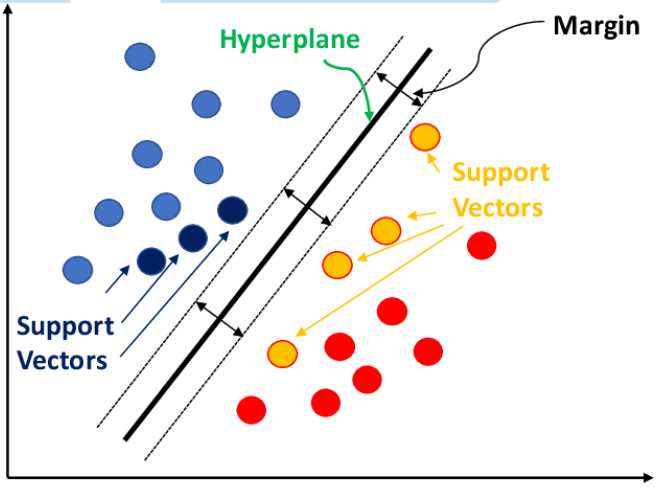
Support Vector Machine digunakan untuk menyelesaikan masalah klasifikasi persamaan linear atau pertidaksamaan linear. Dengan mengubah teknik kernel menjadi *Support Vector Machine* yang akan melakukan *hyperplane* ke dalam dua kelas dalam ruang

vektor, maka masalah pertidaksamaan linear dapat diselesaikan. Berikut ini adalah persamaan tabel dari rumus *kernel* [28]:

Tabel 2. 4 Rumus *Kernel*

Jenis *Kernel*	Model
Linear	$K(x.x') = x.x'$
Polynomial	$K(x.x') = (x.x' + c)'$
RBF Gaussian	$K(x.x') = exp(-y\|\|x-x'\|\|^2)$
Siqmoid	$K(x.x') = tanh(\alpha x.x' + \beta)$

Terlebih lagi, *Support Vector Machine* juga dimanfaatkan untuk mencari *hyperplane* optimal dengan tujuan memaksimalkan jarak antara kelas-kelas yang berbeda. *Hyperplane* sendiri berfungsi sebagai pemisah antar *class*. *support vector* adalah item data terluar dalam *Support Vector Machine* yang paling dekat dengan *hyperplane*.



Gambar 2. 2 *Support Vector Machine Visualization* [29]

Pendekatan yang digunakan Pada Gambar 2.2 [29] merupakan fitur *Support Vector Machine* yang dapat digunakan untuk memisahkan dua *cluster* dan memaksimalkan lebar margin. Dengan memaksimalkan optimasi yang kompleks pada *hyperlane*, maka akan dimanfaatkan pada lebar margin pada setiap titik-titik tersebut. Titik-

titik tersebut akan sangat penting dalam menentukan *hyperplane*, karena akan mendukung setiap margin atau vektor pendukung.

2.2.6 *Logistic Regression*

Logistic Regression (LR) metode *classification text* yang memanfaatkan probabilitas untuk memprediksi karakterisasi informasi yang sebenarnya [30]. Jika fungsi *sigmoid* (nilai output), koneksi linier dapat dibuat antara titik data, dan bilangan koefisien dapat digunakan untuk menentukan hasil. Regresi logistik biasanya dilakukan untuk menghubungkan antara satu atau beberapa variabel independen (variabel bebas) dengan variabel dependen yang biasanya berupa kategori misalnya 0 atau 1.

Tujuan dari metode regresi logistik digunakan untuk menghitung sebuah peluang, melihat karakteristik dan juga melihat faktor- faktor apa saja yang memengaruhi. Ketika ingin melakukan sebuah perhitungan harus menentukan asumsi yang digunakan, karena tidak semua jenis data dapat dianalisis dengan menggunakan regresi logistik.

Spekulasi yang harus dipenuhi dalam sebuah data agar prosedur regresi logistik ini dapat diterapkan dengan tepat, misalnya jika peneliti memiliki keinginan untuk menggunakan prosedur regresi logistik, peneliti tidak perlu menekankan pada jenis hubungan antara faktor bebas dan variabel dependen, hal ini dengan alasan bahwa ketika peneliti harus melakukan sebuah prosedur regresi logistik, maka variabel independen tidak ada hubungannya dengan variabel dependen. Oleh sebab itu, untuk menggunakan regresi logistik, tipe data harus dikotomi, dengan hanya dua kategori. Jika pada analisis dengan metode regresi logistik ini terdapat asumsi *homoskedastisitas* yang harus diperlukan, maka metode regresi logistik tidak harus terpenuhi, dengan kata lain analisis tersebut masih dapat berjalan meskipun datanya tidak *homoskedastisitas*. Fungsi dari

homoskedastisitas dalam analisis statistik yang mengacu pada asumsi bahwa variabilitas (varians) dari kesalahan (residuals) dalam sebuah model regresi atau analisis regresi tidak bergantung pada nilai-nilai prediktor atau variabel independent [31]. Berikut ini merupakan persamaan dari metode regresi logistik [30]:

$$ln\left(\frac{\rho}{1-\rho}\right) = B_0 + B_1 X \quad \textbf{(2.3)}$$

Keterangan:

ln = Logaritma natural

B_0 = Konstanta

B_1 = Koefisien masing-masing variabel

X = Variabel independen

ρ = Probabilitas logistik yang dirumuskan sebagai berikut:

$$\rho = 1 + \frac{e(B_0 + B_1 X)}{1 + e(B_0 + B_1 X)} = \frac{e^{B_0 + B_1 X}}{1 + e^{B_0 + B_1 X}} \quad \textbf{(2.4)}$$

Keterangan:

e atau exp = Fungsi eksponen

Dengan metode regresi logistik tersebut, tentunya akan rumit untuk menginterpretasikan koefisien dari regresinya, sehingga untuk mempermudah mengidentifikasi karakteristik pada suatu data biasanya memakai nilai *odds ratio* atau nilai eksponen dari koefisien regresi. Dengan demikian, dengan memakai nilai *odds ratio* dapat menandakan suatu variabel tersebut.

2.2.7 *Confusion Matrix*

Confusion matrix adalah sebuah tabel yang digunakan untuk menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah [32]. Alat analisis prediktif yang dikenal sebagai

confusion matrix akan menampilkan dan membandingkan nilai aktual (nilai sebenarnya) dengan nilai prediksi dari metode yang digunakan dalam pemrosesan data untuk menghasilkan evaluasi model dalam hal akurasi, presisi*, recall,* dan *f1-score.* Berikut ini merupakan perhitungan untuk mengukur performa dari model klasifikasi untuk dilakukan prediksi:

a. *Accuracy*

Nilai akurasi akan dihitung dengan membagi jumlah total data pada *dataset* (TP + FP + FN + TN) dengan jumlah data yang memiliki nilai positif dan diprediksi sebagai *true positive* (TP) dan jumlah data yang memiliki nilai negatif dan diprediksi sebagai *true negative* (TN) [24].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad \textbf{(2.5)}$$

b. *Precision*

Nilai presisi akan dihitung berdasarkan jumlah data yang memiliki nilai positif dan diprediksi benar positif (TP) dibagi dengan jumlah data yang kategorinya positif (TP + FP) [24].

$$Precision = \frac{TP}{TP + FP} \quad \textbf{(2.6)}$$

c. *Recall*

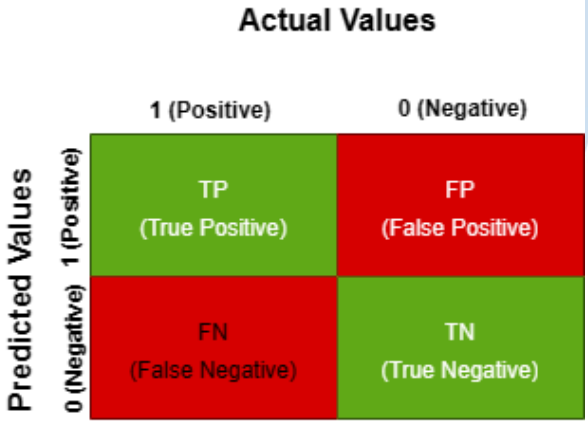
Nilai *recall* akan ditampilkan dengan menggunakan jumlah total data positif dan diprediksi benar positif (TP) dibagi dengan jumlah data yang kategorinya negatif (TP + FN) [24].

$$Recall = \frac{TP}{TP + FN} \quad \textbf{(2.7)}$$

d. *F1-Score*

Nilai *f1-score* akan direpresentasikan dari hasil antara nilai *precision* dan juga nilai *recall* antara kategori yang diprediksi dengan kategori sebenarnya [24].

$$F1 = \frac{2\ x\ Recall\ x\ Precision}{Recall\ +\ Precision} \tag{2.8}$$



Gambar 2. 3 *Confusion Matrix*

Pada Gambar 2.3 menggambarkan empat nilai yang dihasilkan oleh tabel *confusion matrix*, yaitu *true positive* (TP) yang merupakan jumlah data positif yang diprediksi dengan benar sebagai positif, *false positive* (FP) yang merupakan jumlah data negatif yang salah diprediksi sebagai positif, *false negative* (FN) yang merupakan jumlah data positif yang salah diprediksi sebagai negatif, dan *true negative* (TN) yang merupakan jumlah data negatif yang diprediksi dengan benar sebagai negatif.

2.2.8 *Area Under Curve*

Area under curve atau AUC adalah nilai *receiver operating characteristic* (ROC) untuk melakukan evaluasi model klasifikasi untuk menentukan *threshold* (ambang batas). *Thredshold* sendiri berfungsi untuk membedakan antara kelas positif dan kelas negatif

dalam masalah klasifikasi biner [33]. ROC *Curve* yang juga dikenal sebagai kurva *True Positive* (TP) atau *False Positive* (FP) menggambarkan nilai yang menunjukkan nilai negatif meskipun diprediksi positif dan menggambarkan performa dari model klasifikasi tanpa harus memperhatikan kategori kelas atau error. Dengan kata lain, kurva ROC akan menggambarkan nilai yang merupakan jumlah data yang positif dan diprediksi dengan benar [34].

Secara sederhana, ROC dengan nilai antara 0.0 dan 1.0 harus dipertimbangkan ketika menghitung area kurva ROC [30]. Hasilnya, nilai kinerja ROC yang baik dapat ditentukan dengan mengukur luas kurva. Berikut ini merupakan penilai *Area Under Curve*:

Tabel 2. 5 Nilai *AUC*

Nilai AUC	Tingkat Klasifikasi
0.9 – 1.00	Klasifikasi yang luar biasa
0.8 – 0.9	Klasifikasi yang baik
0.7 – 0.8	Klasifikasi yang cukup baik
0.6 – 0.7	Klasifikasi yang buruk
0.5 – 0.6	Klasifikasi yang gagal

2.3 Teori tentang Tools / Software yang digunakan

2.3.1 Twitter

Twitter adalah platform media sosial yang memungkinkan pengguna mengekspresikan pendapat atau komentar mereka dalam bentuk *tweet* [35]. Dalam menggunakan sebuah Twitter, pengguna harus terlebih dahulu melakukan registrasi terlebih dahulu, setelah pengguna berhasil melakukan registrasi, maka pengguna memiliki kemampuan untuk memanfaatkan berbagai fitur yang tersedia di dalam platform media sosial Twitter.

Fasilitas lainnya yang ditawarkan di media sosial Twitter adalah dapat menggunakan *API* Twitter, dimana pengguna memungkinkan untuk mengakses informasi dari Twitter. Ketika ingin mendapatkan

API Twitter harus mendaftar menjadi *developer* Twitter. Setelah itu akan mendapatkan sebuah akses *token* yang digunakan untuk mengakses informasi ke Twitter.

2.3.2 Python

Python merupakan salah satu bahasa pemrograman tingkat tinggi yang berorientasi pada suatu objek dan bersifat *open source*. Bahasa pemrograman Python sangat luas penerapannya seperti bidang dalam pembuatan *website*, mengolah data, bahkan sampai pembuatan *game* [36]. Python juga disebut sebagai bahasa pemrograman yang terdapat berbagai macam *library open source* yang sangat lengkap dan jelas.

Dengan cara ini, Python dianggap mampu dalam menangani pembuatan *Big Data*, *Data Mining*, *Data Science*, *Deep Learning*, bahkan yang lagi banyak yaitu *machine learning* [36]. Oleh sebab itu, Python adalah suatu bahasa pemrograman yang simpel untuk membuat *artificial intelligence*.

2.4 Penelitian Terdahulu

Berikut ini merupakan penelitian beberapa penelitian sebelumnya terkait dengan pinjaman *online* yang akan dijadikan acuan penulis pada penelitian yang akan dilakukan:

Tabel 2. 6 Penelitian Terdahulu

No	Penulis	Tujuan	Metode	Dataset	Hasil Akurasi
1.	Dian Siti Utami, Adhitia, Erfina [1].	Untuk mendapatkan ulasan masyarakat mengenai pinjaman *online*	*SVM*	Data Twitter Pinjaman *Online*	62%
2.	Tri Puji Lestari [3].	Untuk melihat sentimen mengenai pinjaman *online*	*SVM* dan *Social Network Analysis*	Data Twitter Pinjaman *Online*	86.6%

No	Penulis	Tujuan	Metode	Dataset	Hasil Akurasi
3.	Sri Handika Utami, Anton Ade Purnama, Achmad Nizar Hidayanto [37].	Untuk memeriksa kepercayaan terhadap *Fintech Lending* di Indonesia	*KNN*, *SGD*, *RF*, *NN*, *NB*, *LR*, *AdaBoost*	Data Twitter Pinjaman *Online*	*KNN*: 60.6%, *SGD*: 66.7%, *RF*: 75.4%, *NN*: 71.2%, *NB*: 76%, *LR*: 74.3%, *AdaBoost*: 67.7%
4.	Beibei Niu, Jinzheng Ren, Ansa Zhao, Xiaotao Li [38].	Untuk memeriksa kepercayaan terhadap *P2P Lending*	*Latent Dirichlet Allocation*	Wdzj.com	77.45%
5.	Ryan Randy Suryono, Indra Budi [39].	Untuk melihat sentimen pada berita *online* Indonesia	*SVM*, *MNB*, *LR*, dan *RF*	Artikel portal berita *online* Indonesia	*MNB*: 62.14%, *LR*: 62.62%, SVM: 63.61%, *RF*: 61.34%
6.	Kelvin, Jepri Banjarnahor, Evta Indra, Stiven Hamonangan Sinurat [40].	Untuk melihat sentimen mengenai *COVID-19*	*LR*, *SVM*	Data Twitter *COVID-19*	*LR*: 87.68%, *SUPPORT VECTOR MACHINE*: 91.15%
7.	Ni Luh Putu Chandra Savitri, Radya Amirur Rahman, Reyhan Venyutzky, Nur Aini Rakhmawati [9].	Untuk menetapkan kebijakan pemerintah terkait sekolah daring	*LR*, *SVM*, *BNB*, *RF*	Data Twitter Sekolah Daring	*LR*: 88%, *SVM*: 87%, *BNB*: 74%, *RF*: 87%
8.	Dinar Ajeng Kristiyanti, Akhmad Hairul Umam,	Untuk melihat sentimen mengenai pasangan	*SVM, NB*	Data Twitter dengan kata kunci Rindu,	*NB:* 94%, *SVM*: 75.50%

No	Penulis	Tujuan	Metode	Dataset	Hasil Akurasi
	Mochamad Wahyudi, Ruhul Amin, Linda Marlinda [41]	calon gubernur Jawa Barat periode 2018-2023		Hasanah, Asyik, 2DM	
9.	Nicholas, Rudi Sutomo [42]	Untuk mengevaluasi sentimen terhadap *cryptocurrency* dengan pengguna Twitter	*SVM*, *NB*	Data Twitter mengenai *cryptocurrency*	*SVM*: 78.59%, *NB*: 83.81%

Berdasarkan dari hasil jurnal pada Tabel 2.3 yang telah dilampirkan dapat disimpulkan bahwa pada jurnal 1 dan 2 merupakan acuan utama penulis untuk melakukan penelitian ini, karena topik yang diangkat mengenai pinjaman *online*. Selain itu, implementasi mengenai topik yang diangkat juga memiliki metode yang sama digunakan pada jurnal 1 dan 2 menggunakan *Support Vector Machine*. Pada jurnal ke-3 sampai dengan ke-5 akan mengacu kepada ulasan masyarakat mengenai pinjaman *online* dengan target klasifikasi positif dan negatif.

Selain itu, pada jurnal ke-6 dan ke-9 yang telah dilampirkan merupakan salah satu acuan untuk menggunakan metode *Support Vector Machine* dan juga *Logistic Regression* mengenai analisis sentimen. Berdasarkan hasil dari beberapa jurnal yang sudah dilampirkan terbukti memberikan hasil yang baik dibandingkan dengan metode yang lainnya. Oleh karena itu, membandingkan penelitian penulis dengan penelitian sebelumnya pada jurnal 1 dan 2 akan memperoleh tingkat akurasi yang dicapai dengan menerapkan metode *Support Vector Machine* dan membandingkannya dengan metode lain yaitu *Logistic Regression*. Penelitian yang dilakukan penulis akan membandingkan dengan penelitian sebelumnya dalam jurnal 1 dan 2 yaitu hasil evaluasi model dengan menggunakan *confusion matrix* yang akan menampilkan nilai *precision*, *recall*, dan juga *f1-score*. Selain itu, penulis juga akan menambahkan performa model

dengan menggunakan *Area Under Curve* yang sebelumnya tidak ada dalam acuan utama penelitian pada jurnal 1 dan 2.