BAB I
PENDAHULUAN

1.1 Latar Belakang

Pasar modal merupakan salah satu penggerak perekonomian suatu negara, karena pasar modal merupakan tempat pembentuk modal dan akumulasi dana jangka panjang selain itu juga merepresentasikan kondisi perusahaan yang berada di suatu negara. Pasar modal bisa mengalami peningkatan (*bullish*) atau mengalami penurunan (*bearish*) yang dapat dilihat dari naik turunnya harga saham yang tercermin melalui pergerakan indeks atau lebih dikenal dengan Indeks Harga Saham Gabungan (IHSG) (Alfira et al., 2021). Indeks Harga Saham Gabungan (IHSG) merupakan indikator utama kepercayaan investor terhadap pasar modal Indonesia dan berperan penting dalam mencerminkan stabilitas ekonomi nasional. Penurunan IHSG dapat berdampak luas, mulai dari kesulitan perusahaan dalam memperoleh pendanaan, meningkatnya *capital outflow*, hingga pelemahan nilai tukar rupiah yang pada akhirnya berpengaruh terhadap daya beli masyarakat dan pertumbuhan ekonomi.

Pelemahan IHSG baru-baru ini dipicu oleh kombinasi faktor global dan domestik, seperti penurunan peringkat investasi oleh Goldman Sachs dan isu pergantian Menteri Keuangan. Ketidakpastian ini memperburuk sentimen pasar dan mendorong aksi jual, terutama oleh investor asing (ITS News, 2025). Dari Maret hingga April 2025, IHSG tercatat melemah dari kisaran 6.510 poin menjadi sekitar 6.000 poin, atau turun sekitar 7,9%, yang mencerminkan menurunnya kepercayaan investor terhadap stabilitas pasar modal Indonesia (Infobanknews, 2025).

Di tengah dinamika pasar yang penuh ketidakpastian, media sosial telah menjadi salah satu sumber informasi dan opini publik yang paling aktif dan *real-time*. Twitter, yang kini dikenal sebagai X, telah berkembang menjadi *platform* signifikan dalam menyuarakan pendapat masyarakat terhadap isu-isu ekonomi, termasuk fluktuasi IHSG. Dengan lebih dari 140 juta pengguna aktif dan ratusan juta *tweet* setiap harinya (Jhosefhin & Dewi 2025), X menyediakan akses cepat terhadap sentimen dan opini publik yang kaya dan beragam (Wahyudi et al., 2024). *Platform* ini memainkan peran penting dalam membentuk persepsi masyarakat terhadap stabilitas ekonomi dan pasar modal.

Seiring dengan perkembangan teknologi, analisis sentimen berbasis *machine learning* semakin banyak dimanfaatkan untuk mengidentifikasi dan ini masyarakat di media sosial. Berbagai algoritma *machine* ort Vector Machine (SVM) dan *Multinomial Naïve Bayes* dan elah digunakan untuk mengklasifikasikan sentimen dalam teks. iliki keunggulan dan keterbatasan masing-masing, bergantung lan metode praproses yang diterapkan. Misalnya, Algoritma *ine* dikenal dengan kemampuannya dalam klasifikasi data yang

sangat baik, terutama untuk *dataset* yang besar dan memiliki dimensi tinggi (Ikhsani et al., 2024), sementara *Multinomial Naïve Bayes* cocok digunakan dalam klasifikasi fitur diskrit (seperti, jumlah kata dalam klasifikasi teks) (Toyibah et al., 2024).

Berbagai penelitian terdahulu telah membuktikan bahwa media sosial, khususnya X, dapat merefleksikan dinamika sentimen publik secara aktual dan berdampak pada keputusan ekonomi. Namun, belum banyak penelitian yang secara khusus mengaitkan analisis sentimen X dengan isu pelemahan IHSG dan dampaknya terhadap stabilitas ekonomi Indonesia.

Berdasarkan uraian di atas, peneliti tertarik untuk melakukan penelitian yang berjudul "**Analisis Sentimen Aplikasi X terhadap Melemahnya IHSG Menggunakan Metode *Support Vector Machine (SVM)* dan *Multinomial Naïve Bayes***". Penelitian ini diharapkan dapat berkontribusi dalam memahami persepsi publik terhadap kondisi pasar modal Indonesia, serta menyediakan *insight* yang relevan bagi investor, pelaku pasar, dan pembuat kebijakan dalam menghadapi dinamika ekonomi nasional.

1.2 Rumusan Masalah

Rumusan masalah yang dibahas pada penelitian ini adalah:

Bagaimana performa model *Support Vector Machine* (SVM) dan *Multinomial Naïve Bayes* dalam mengklasifikasikan sentimen pengguna Aplikasi X terkait melemahnya IHSG?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah memperoleh hasil performa model *Support Vector Machine* (SVM) dan *Multinomial Naïve Bayes* dalam mengklasifikasikan sentimen pengguna Aplikasi X terhadap melemahnya IHSG.

1.4 Landasan Teori

1.4.1 Analisis Sentimen

Analisis sentimen adalah sebuah teknik atau proses yang digunakan untuk mengevaluasi dan mengidentifikasi sentimen, perasaan, atau emosi yang terkandung dalam teks, seperti dokumen, artikel, ulasan, atau pesan media sosial.

Tujuan dari analisis sentimen adalah untuk memahami dan mengukur reaksi
apat orang terhadap suatu topik, produk, layanan, atau peristiwa
telah terkumpul akan diolah melalui analisis agar dapat
gan yang mendalam sehingga kesimpulan dapat diambil. Hasil
at berupa sentimen positif, negatif, atau netral.

Analisis Sentimen dapat dibagi menjadi 3 level, yaitu:

a. Level Dokumen: Analisis sentimen bertujuan untuk mengklasifikasikan keseluruhan dokumen ke dalam kelas positif, netral, atau negatif.
b. Level Kalimat: Analisis sentimen bertujuan untuk menentukan sentimen positif atau negatif dari suatu kalimat dengan mempertimbangkan susunan kata dalam kalimat tersebut.
c. Level Aspek: Analisis sentimen bertujuan untuk menentukan sentimen positif, negatif, atau netral berdasarkan atribut dari suatu entitas (Manullang et al., 2023).

Proses analisis sentimen dipengaruhi oleh *dataset* yang digunakan. Untuk *dataset* yang terdiri dari kumpulan kalimat yang cukup panjang membutuhkan penanganan yang berbeda (widayat, W., 2021). Analisis sentimen adalah salah satu *domain* dari *Natural Language Processing* (NLP) (Ndapamuri, 2023).

Natural Language Processing (NLP) adalah cabang ilmu komputer dan kecerdasan buatan yang berfokus pada pemahaman, pemrosesan, dan generasi bahasa manusia oleh komputer. NLP memungkinkan komputer untuk berinteraksi dengan manusia melalui bahasa alami, yang dapat berupa ucapan atau teks. Tujuan utama NLP adalah untuk memahami, menganalisis, dan menghasilkan teks dalam bahasa manusia, serta mengembangkan sistem yang dapat memproses dan berkomunikasi dengan manusia dengan cara yang mirip dengan komunikasi antarmanusia (Rivaldi & Wismarini, 2024).

1.4.2 Aplikasi X

Twitter atau sekarang disebut X merupakan salah satu dari berbagai macam *platform* media sosial yang biasanya digunakan untuk berbagi opini kepada masyarakat, terdapat berbagai fitur seperti mengambil gambar, membagikan, serta mengomentari postingan dengan konteks tertentu (Nopan & Mailoa, 2024). Twitter digunakan oleh berbagai kalangan, termasuk individu, perusahaan, organisasi, selebriti, dan politisi. Twitter juga mengumumkan bahwa mereka memiliki lebih dari 140 juta pengguna aktif yang membuat 340 juta *tweet* tiap hari sehingga tidak heran jika Twitter sering digunakan untuk berbagai berita terkini, berkomunikasi, mempromosikan produk atau layanan (Jhosefhin & Dewi, 2025).

Twitter telah menjadi *platform* signifikan dalam pencarian informasi aktual, menawarkan beragam fitur yang memenuhi kebutuhan penggunanya. Di era kemajuan teknologi yang pesat ini, media sosial, khususnya Twitter, telah menjadi kehidupan sehari-hari masyarakat. *Platform* ini tidak hanya berfungsi sebagai sarana penting bagi publik untuk berbagi ekspresikan pendapat melalui *tweet.* Keunggulan Twitter terletak menyediakan akses cepat ke opini pengguna, menjadikannya g kaya dan beragam, mencerminkan pandangan dan perasaan ntang berbagai topik terkini (Wahyudi et al., 2024).

1.4.3 Indeks Harga Saham Gabungan (IHSG)

Indeks Harga Saham Gabungan (IHSG) merupakan indeks gabungan dari seluruh jenis saham yang tercatat di Bursa Efek Indonesia (BEI). Indeks Harga Saham Gabungan merupakan permulaan pertimbangan untuk melakukan investasi. Hal ini disebabkan, Indeks Harga Saham Gabungan dapat digunakan untuk menilai situasi pasar secara umum atau mengukur apakah harga saham mengalami kenaikan atau penurunan (Jaya et al., 2023).

Indeks Harga Saham Gabungan dianggap sebagai dasar analisis yang paling sering digunakan oleh para analis untuk melihat kondisi saham di pasar modal Indonesia. Semakin berkembang pasar modal nasional maka IHSG akan semakin meningkat. Peningkatan atas pasar modal (*bullish*) atau penurunan (*bearish*) dapat ditinjau dari kenaikan dan penurunan keseluruhan harga saham yang tampak melalui pergerakan (Mujaddidi & Adnan, 2024).

1.4.4 *Crawling*

Crawling merupakan teknik yang digunakan dalam pengumpulan data dari berbagai jenis sumber baik berupa *website* atau media sosial (Nopan & Mailoa, 2024). Data *crawling* adalah teknik otomatis untuk mengumpulkan dan mengindeks data dari berbagai sumber *online*, menggunakan program yang disebut "*crawler*" atau "*spider*" untuk menjelajahi halaman *web*, mengikuti tautan, dan mengumpulkan data. Data yang diperoleh bisa digunakan untuk analisis data, penelitian, atau pengembangan sistem informasi.

Data crawling dan *web crawling* mempunyai tujuan yang sama yaitu mengumpulkan data dari sumber *online*, namun perbedaan dari keduanya:

1. Sumber data : *web crawling* lebih fokus pada data dari internet seperti situs web, sedangkan *data crawling* dapat melibatkan data dari *database,* file, hingga API.
2. Skala : *web crawling* biasanya dilakukan pada skala besar, sementara *data crawling* bisa skala kecil maupun besar tergantung kebutuhan.

Data *Crawling* dapat diotomatisasi sebagai bot atau robot untuk mengambil data dari halaman *web* tertentu (Talisman et al., 2024).

1.4.5 *Text Preprocessing*

Text Preprocessing merupakan tahap yang penting sebelum melakukan sentimen analisis untuk mendapatkan hasil akurasi yang lebih tinggi (Oktaviana et al., 2022).

pakan tahap dalam proses membersihkan data yang akan
bertujuan untuk mengubah data bentuk mentah menjadi data
a data yang telah siap bisa digunakan pada proses klasifikasi
ma yang sudah ditentukan (Nopan & Mailoa, 2024).
t penting dalam analisis sentimen, terutama dalam sebuah

platform yang sebagian besar berisi kata atau kalimat yang tidak formal, tidak terstruktur, serta memiliki *noise* yang cukup besar (Ratiasasadara et al., 2023).

Tahapan *preprocessing* yang dilakukan adalah sebagai berikut:

a. *Cleaning*
   Tahap pembersihan (*cleaning*) bertujuan untuk menghilangkan atau mengurangi kata maupun kalimat yang tidak relevan dalam data *tweet*, seperti tanda baca, karakter *unicode*, dan elemen lainnya. Proses *cleaning* ini terdiri dari lima tahapan yang akan dijalankan oleh sistem guna memperoleh hasil yang optimal, di antaranya: Membersihkan tanda baca; Membersihkan angka; Membersihkan *link*, Membersihkan kelebihan spasi, *URLs*, *Hashtags*, *Mentions*, *Reservedwords* (RT,FAV), *Emojis*, *Smileys*.
b. *Case Folding*
   Pada tahap ini, seluruh huruf dalam teks akan dikonversi menjadi huruf kecil (l*owercase*). Adapun proses *case folding* meliputi langkah-langkah berikut: memeriksa setiap karakter dari awal hingga akhir teks, lalu mengubah huruf kapital (*uppercase*) menjadi huruf kecil (*lowercase*) jika ditemukan (Prasetyo et al., 2025).
c. *Normalization*
   Teks yang tidak terstruktur memerlukan proses transformasi menjadi bentuk tulisan yang teratur untuk mempersingkat proses eksplorasi data dalam analisis sentimen. Pada langkah ini dilakukan normalisasi teks pada data frame untuk membakukan kata singkatan, misalnya mengubah "yg" dijadikan "yang", "dgn" dirubah "dengan", "udh" dijadikan "udah", "tp“ dijadikan "tetapi", dll. Selain itu, kata-kata seperti "udah" dan "sudah", serta “tidak" dan "nggak", dianggap memiliki arti yang sama (Putri et al., 2025).
d. *Tokenizing*
   Tahap *tokenizing* berfungsi untuk memisahkan kata-kata dalam *tweet* menjadi unit-unit yang lebih kecil. Proses ini menggunakan spasi sebagai pemisah antar kata. Dalam sebuah *tweet*, terdapat sejumlah kata yang terhubung dan dipisahkan oleh spasi. Agar teks lebih mudah diproses, setiap kata dalam kalimat harus dipisahkan. Jika suatu karakter bukan tanda pemisah seperti titik (.), koma (,), atau spasi, maka karakter tersebut akan digabungkan dengan karakter berikutnya.
e. *Stopwords Removal*
   Pada tahap ini, kumpulan *tweet* yang telah melewati proses *tokenizing* akan diproses lebih lanjut dengan menghapus kata-kata yang tidak memiliki makna analisis sentimen. Setiap kata dalam *tweet* akan diperiksa, dan ut termasuk dalam kategori kata sambung, kata depan, kata yang tidak relevan, maka kata tersebut akan dihapus.

um *tweet* seringkali memiliki berbagai bentuk morfologi. Oleh ap kata akan diubah menjadi bentuk dasarnya (*stem*) yang

sesuai. Kata-kata ini akan direduksi dengan menghilangkan awalan atau akhiran yang ada (Prasetyo et al., 2025). *Stemming* dibutuhkan untuk meminimalisir indeks yang berbeda dengan menjadikan suatu kata yang memiliki makna yang sama menjadi bentuk yang sama juga, seperti “terbukti” dan “membuktikan” menjadi “bukti”, kemudian “dilonggarkan” dan “pelonggaran” menjadi “longgar”, dan sebagainya (Ratiasasadara et al., 2022).

1.4.6 *Lexicon Based*

Lexicon-Based (berbasis leksikon) merupakan salah satu metode dalam analisis sentimen yang menggunakan kamus leksikon yang telah dikurasi secara manual. Kamus leksikon ini berisi daftar kata-kata atau frasa yang memiliki hubungan dengan sentimen tertentu, seperti positif, negatif, atau netral. Setiap kata atau frase dalam kamus diberi label sentimen yang sesuai*. Lexicon Based* adalah fitur kata yang memiliki sentimen positif atau negatif berdasarkan kamus atau *lexicon*. Proses pelabelan data dilakukan oleh kamus *Lexicon Based* dengan menghitung skor sentimen. Setelah kata-kata yang mengandung sentimen positif, negatif, dan netral diidentifikasi dalam sebuah kalimat, langkah selanjutnya adalah menghitung setiap kata yang mengandung sentimen dalam kalimat tersebut, dengan menjumlahkan nilai opini. Jumlah nilai opini untuk sentimen positif bernilai 1 atau lebih, sedangkan kata-kata yang netral dalam kalimat diberi nilai = 0, dan sebaliknya, untuk kata-kata yang mengandung sentimen negatif dalam kalimat diberi nilai = -1 atau lebih (Manullang et al., 2023).

1.4.7 TF-IDF

Feature Extraction (Fitur Ekstraksi) merupakan tahap yang paling penting sebelum proses klasifikasi. Tujuan dari tahapan ini adalah untuk mengahasilkan sekumpulan fitur dengan melakukan pembobotan kata, pembobotan kata tersebut dapat dihitung sehingga dapat mengklasifikasikan sebuah data. TF-IDF merupakan metode yang digunakan untuk memberikan bobot pada *term* sebagai strategi untuk mengklasifikasikan dokumen. Proses pembobotan pada *term* menggunakan TF-IDF terdiri dari menghitung nilai TF (*Term Frequency*) yaitu untuk menghitung suatu kata yang diulang dalam sebuah teks, sementara IDF (*Invers Document Frequency*) yaitu untuk menghitung probabilitas dalam suatu kata pada teks (Wicaksono et al., 2023).

Term frequency menyatakan nilai frekuensi *term* yang sering muncul pada sebuah dokumen. Semakin besar jumlah kemunculan suatu *term* dalam dokumen, semakin besar pula bobotnya atau akan memberikan nilai kesesuaian yang semakin besar. Menurut Wibowo et al., 2024, rumus TF dapat didefinisikan sebagai berikut:

$$TF(t,d) = \frac{f(t,d)}{\sum_k f(k,d)} \quad (1)$$

kuensi kemunculan term t pada dokumen d

n (Frasa/kata)

d = Dokumen
$\sum_k f(k,d)$ = Total seluruh term dalam dokumen d

Selanjutnya, IDF dihitung dengan penambahan *Smoothing*, yakni menambahkan angka 1 pada pembilang dan penyebut untuk mencegah pembagian dengan nol (Ratiasasadara et al., 2022). Berikut rumus untuk menghitung IDF yang ditunjukan pada persamaan (2):

$$IDF\ (t) = \log\left(\frac{N+1}{df(t)+1}\right) + 1 \quad (2)$$

Keterangan:

N = Jumlah dokumen pada dataset
$df(t)$ = Jumlah dokumen yang mengandung *term* t

TF-IDF dapat dirumuskan sebagai berikut:

$$TF\ IDF(t,d) = TF \times IDF \quad (3)$$

Keterangan:

TF = Nilai dari *term frequency*
IDF = Nilai dari *inverse document frequency*

Sehingga, perhitungan (*Term frequency-inverse document frequency*) TF-IDF dengan modul *scikit-learn* dapat dihitung dengan menggunakan dengan Persamaan (4):

$$W_{t,d} = \frac{tfidf_{t,d}}{\sqrt{\sum_{d=1}^{n} tfidf_{t,d}^2}} \quad (4)$$

Keterangan:

$W_{t,d}$ = Bobot akhir term t pada dokumen d
$tfidf_{t,d}$ = Bobot $TF\ IDF$ term t pada dokumen d

1.4.8 *Support Vector Machine*

Menurut Wahyudi et al., (2024) *Support Vector Machine* (SVM) adalah sebuah basis *machine learning*, memprediksi kelas berdasarkan model lkan dari proses pelatihan. SVM mempunyai keunggulan yang mbelajaran statistika dan performanya yang sering melampaui klasifikasi SVM melibatkan tahap pelatihan yang menghasilkan , dilanjutkan dengan tahap pengujian. Pada tahap pengujian, la yang telah dipelajari untuk memberikan label sentimen pada

tweet, sekaligus menghasilkan tingkat akurasi klasifikasi. Dengan demikian, SVM tidak hanya mampu mengklasifikasikan data baru, tetapi juga memberikan ukuran kehandalan hasil klasifikasinya. Berikut merupakan rumus umum *Support Vector Machine* (SVM):

$$f(x) = sign\,(w \cdot x + b) \tag{5}$$

Keterangan:

$f(x)$ = Fungsi prediksi yang menentukan label kelas dari input x
$sign$ = Fungsi tanda yang menghasilkan *output* +1 jika argumen positif dan -1 jika argumen negatif. Dalam konteks SVM, fungsi ini digunakan untuk menentukan ke kelas mana data x akan diklasifikasikan.
w = Vektor bobot yang menentukan arah dan orientasi *hyperplane.*
x = Merupakan vektor fitur dari data *input*
b = Bias atau *intercept* dari *hyperplane*

1.4.9 *Multinomial Naïve Bayes*

Multinomial Naïve Bayes merupakan pengembangan dari N*aïve Bayes Classifier* yang cocok digunakan untuk klasifikasi teks (Ratiasasadara et al., 2022). Menurut Lundam & Darmawan 2024, *Multinomial Naïve Bayes* adalah metode klasifikasi yang menggunakan probabilitas yang sangat efektif untuk data teks dan frekuensi kejadian. Metode ini mengasumsikan independensi kuat antara fitur dan bekerja berdasarkan teorema *Bayes*, dengan tahapan menghitung jumlah kelas atau label, menghitung jumlah kelas yang sesuai dengan kasus yang diamati, mengalikan variabel-variabel kelas, dan membandingkan hasilnya untuk menentukan kelas yang paling mungkin. Berikut rumus untuk menghitung *Multinomial Naïve Bayes* yang ditunjukan pada persamaan (6).

$$P(H|X) = \frac{P(X|H)}{P(X)} P(H) \tag{6}$$

Keterangan:

X = Data dengan kelas yang belum diketahui
H = Hipotesis data merupakan suatu kelas spesifik
$P(H|X)$ = Probabilitas hipotesis H berdasarkan kondisi X (*Posterior Probabilitas*)
$P(H)$ = Probabilitas hipotesis H (Prior probabilitas)
$P(X|H)$ = Probabilitas X berdasarkan kondisi pada hipotesis
s X.

ndingkan antara kelas untuk menentukan klasifikasi. Rumus
ditunjukan pada persamaan (7).

$$P(C|d) = \frac{P(C) \cdot P(d|C)}{P(d)} \tag{7}$$

Keterangan :

$P(C|d)$ = Probabilitas sampel d dalam kelas C
$P(C)$ = *Probabilitas prior* kelas C
$P(d|C)$ = *Likelihood*, probabilitas sampel d dalam kelas C
$P(d)$ = *Evidence*, probabilitas keseluruhan dari sampel d muncul diseluruh kelas

Untuk perhitungan *likelihood* dihitung menggunakan persamaan (8).

$$P(d|C) = \prod_{i=1}^{n} p(x_i|C) f_i \quad (8)$$

Keterangan :

n = Jumlah fitur dalam sampel d.
x_i = Kata ke-i dalam sampel d.
f_i = Jumlah kemunculan fitur x_i dalam sampel d
$p(x_i|C)$ = Probabilitas fitur x_i muncul dalam kelas C.

1.4.10 *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) digunakan untuk menangani *dataset* yang tidak seimbang dengan menambah data sintetis pada kelas minoritas agar seimbang dengan kelas mayoritas. Metode ini diterapkan karena ketidakseimbangan antara jumlah data sentimen positif, netral, dan negatif. *Oversampling* dilakukan untuk menyeimbangkan jumlah data dengan menambahkan kelas yang lebih kecil (Lundam & Darmawan, 2024). Tidak memperhatikan ketidakseimbangan ini dapat menyebabkan model memiliki bias yang sangat besar terhadap kelas mayoritas, yang dapat membuat model tidak sensitif terhadap hal-hal yang berkaitan dengan kelas minoritas yang mungkin memiliki nilai prediktif yang signifikan. Karena kelas mayoritas berkuasa, akurasi model dapat sangat tinggi, tetapi hasil prediksi untuk kelas minoritas akan sangat rendah (Wahyudi et al., 2024).

1.4.11 *Confusion Matrix*

Confusion Matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja atau performa dalam permasalahan klasifikasi biner maupun permasalahan klasifikasi *multiclass*. Adapun beberapa kegunaan dari *confusion matrix* diantaranya digunakan untuk mengevaluasi model klasifikasi agar memahami seberapa baik model yang digunakan dalam mengklasifikasikan atau memprediksi data, digunakan untuk melakukan perbandingan kinerja beberapa model yang k mengidentifikasi jenis kesalahan yang sering terjadi dan Stefanni et al., 2025). Terdapat 4 istilah dalam proses klasifikasi

Tabel 1 *Confusion Matrix*

	Prediksi Negatif	Prediksi Positif
Aktual Negatif	*True Negatif* (TN)	*False Positif* (FP)
Aktual Positif	*False Negatif* (FN)	*True Positif* (TP)

Adapun perhitungan performa *Confusion Matrix* sebagai berikut:

1) *Accuracy*
Accuracy digunakan untuk mengukur tingkat kedekatan antara nilai prediksi dan nilai faktual. Rumus perhitungan nilai *Accuracy* dapat dilihat pada persamaan berikut:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (9)$$

2) *Precision*
Precision digunakan untuk mengukur tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Rumus perhitungan *Precision* dapat dilihat pada persamaan 10.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

3) *Recall*
Recall digunakan untuk mengukur tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Rumus perhitungan *Recall* terlihat pada persamaan sibawah ini :

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

4) *F1-Score,*
F1-Score digunakan untuk mengukur keseimbangan antara kemampuan model dalam mengenali ulasan positif dan negatif. Rumus perhitungan *F1-Score* dapat dilihat pada persamaan 12.

$$F1 - Score = \frac{2 \times (Presisi \times Recall)}{(Presisi + Recall)} \quad (12)$$

ıg benar-benar positif
ıg benar-benar negatif
gatif yang diprediksi positif
sitif yang diprediksi negatif

BAB II
METODOLOGI PENELITIAN

2.1 Pendekatan dan Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode komputasi tekstual berbasis *machine learning*. Jenis penelitian ini adalah penelitian terapan (*applied research*) yang bertujuan menguji dan membandingkan performa *Support Vector Machine* (SVM) dan *Multinomial Naïve Bayes* dalam mengklasifikasikan sentimen *tweet* terkait pelemahan Indeks Harga Saham Gabungan (IHSG), serta menghasilkan model analisis sentimen yang dapat dimanfaatkan sebagai bahan pertimbangan keputusan pasar modal.

2.2 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan dalam kurun waktu April hingga Juni 2025. Seluruh kegiatan penelitian, yang meliputi pengumpulan data, pra-pemrosesan data, penerapan algoritma, evaluasi hasil, hingga penyusunan laporan, dilakukan di Laboratorium *Big Data*, Program Studi Ilmu Aktuaria, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin.

2.3 Objek Penelitian

Objek dalam penelitian ini adalah data teks yang berasal dari Aplikasi X yang membahas mengenai melemahnya Indeks Harga Saham Gabungan (IHSG). Data tersebut berisi opini, komentar, atau pernyataan dari pengguna yang dikumpulkan dalam bentuk teks. Data teks ini dianalisis untuk mengetahui sentimennya, apakah bersifat positif atau negatif, dengan menggunakan dua algoritma klasifikasi, yaitu *Support Vector Machine* (SVM) dan *Multinomial Naïve Bayes*. Oleh karena itu, objek penelitian ini mencakup teks dari Aplikasi X serta hasil analisis klasifikasi sentimen menggunakan kedua metode tersebut.

2.4 Metode Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan melalui proses *web crawling* terhadap media sosial X menggunakan alat bantu *Tweet Harvest* versi terbaru. Proses *crawling* dijalankan melalui *Google Colaboratory* dengan bahasa pemrograman *Python*, serta membutuhkan instalasi tambahan berupa *Node.js* karena *Tweet Harvest* dibangun menggunakan *platform* tersebut. Data dikumpulkan berdasarkan kata kunci “IHSG” dengan rentang waktu 27 Maret hingga 3 April 2025 dan menggunakan filter bahasa Indonesia. Pemilihan periode tersebut didasarkan enurunan signifikan IHSG yang disertai meningkatnya di media sosial terkait kondisi pasar modal nasional. Hasil dari enghasilkan sebanyak 1113 *tweet* yang disimpan dalam format a digunakan dalam tahap pra-pemrosesan untuk analisis

Langkah-langkah pengumpulan (*Crawling*) data dilakukan sebagai berikut:

1. Membuat akun X aktif yang diperlukan sebagai syarat untuk mendapatkan *auth token* yang digunakan dalam proses pengambilan data.
2. Mengakses akun X melalui *browser*, lalu mendapatkan *auth token* dengan membuka fitur *Inspect Element*, menuju *tab Application*, memilih bagian *Cookies,* dan menyalin nilai dari *auth_token*. *Token* ini bersifat pribadi dan tidak boleh dibagikan, karena berkaitan langsung dengan akses terhadap akun X.
3. Menginstal *Node.js* di *Google Colab* , karena *Tweet Harvest* dibangun menggunakan teknologi *Node.js*, sehingga *Node.js* perlu terlebih dahulu dilakukan instalasi di *Google Colab* sebelum alat tersebut dapat dijalankan.
4. Menjalankan *Tweet Harvest* dan disertai dengan parameter penting, seperti nama file *output*, kata kunci pencarian, rentang waktu pengambilan data, bahasa, serta jumlah maksimum *tweet* yang diambil.
5. Melakukan proses *crawling* secara otomatis menggunakan *headless browser* yang dijalankan oleh *Tweet Harvest* untuk mengakses X dan mengumpulkan data sesuai parameter.
6. Menyimpan hasil pengambilan data dalam format *file* CSV (*Comma Separated Values*), dengan nama “ihsg.csv” di Google Colab, yang berisi teks *tweet* dan atribut data untuk keperluan analisis selanjutnya.

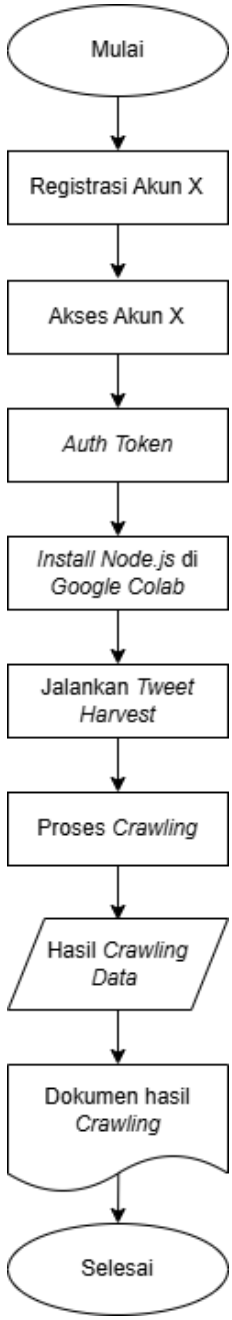
2.5 Metode Analisis Data

Tahapan analisis sebagai berikut:

1. Melakukan eksplorasi data awal (EDA) untuk memahami karakteristik dasar data, termasuk distribusi jumlah tweet dan analisis kata dominan melalui *wordcloud* dan frekuensi katal.
2. Melakukan pra-pemrosesan (*Preprocessing*) data untuk membersihkan dan menyiapkan teks. Tahapan meliputi: *cleaning*, *case folding*, *tokenizing*, *stopword removal*, *normalization*, dan *stemming*.
3. Melabeli data menggunakan pendekatan *lexicon-based* dengan kamus sentimen Bahasa Indonesia dalam skema klasifikasi biner (positif dan negatif).
4. Melakukan ekstraksi fitur menggunakan metode TF-IDF untuk mengubah teks menjadi representasi numerik yang siap digunakan oleh model klasifikasi.
5. Melatih dan menguji model dengan membagi data menjadi data latih (80%) dan data uji (20%), menggunakan algoritma SVM dan *Multinomial Naïve Bayes*, baik sebelum maupun sesudah penyeimbangan data dengan STEMO.
6. Melakukan evaluasi model menggunakan *confusion matrix* serta pengukuran rti *Accuracy*, *Precision*, *Recall*, dan *F1-Score* untuk performa kedua model serta menilai dampak penyeimbangan asil klasifikasi.

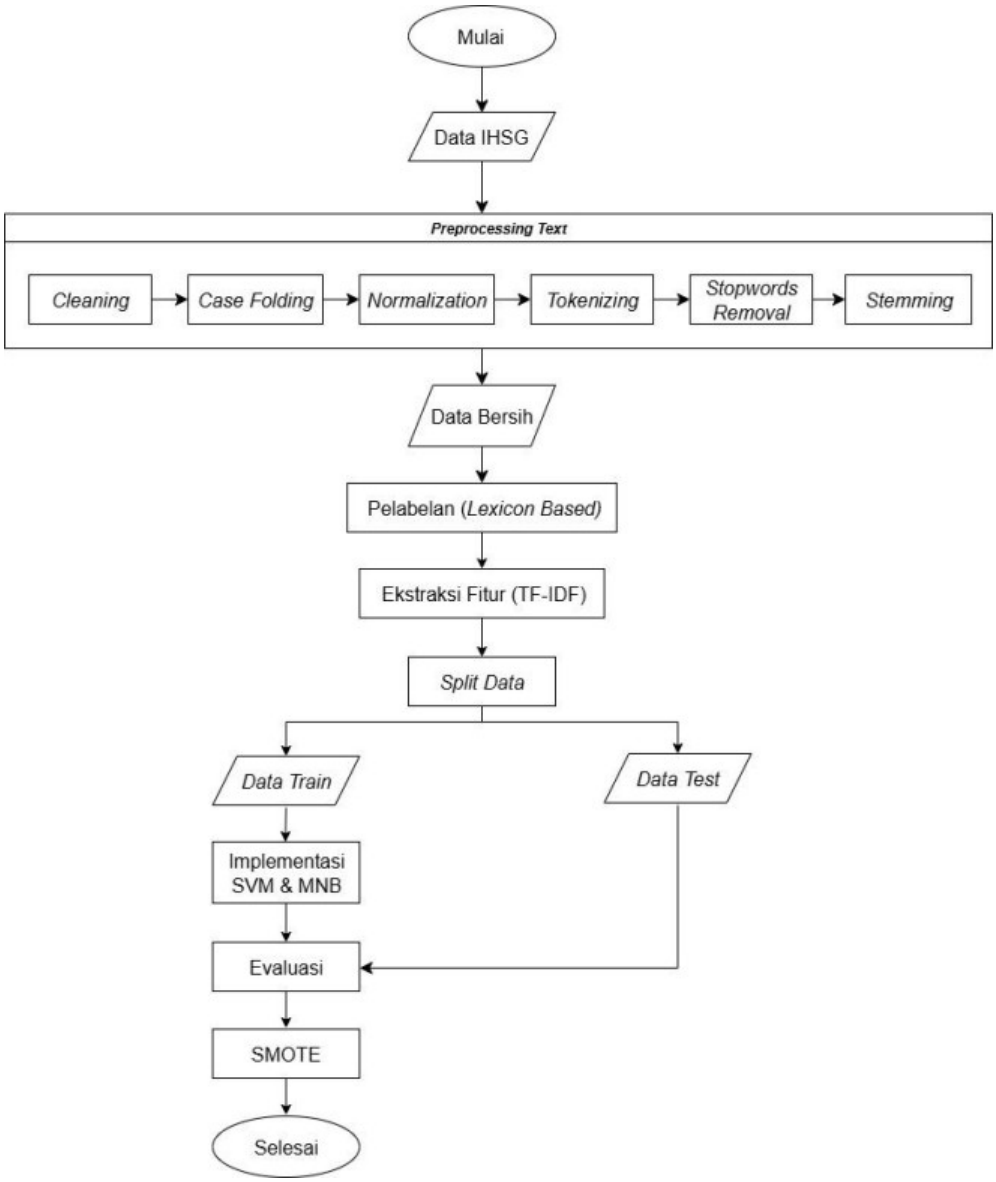
2.6 Alur kerja

2.6.1 Tahapan Pengumpulan Data



pengumpulan data

2.6.2 Tahapan Analisis



Gambar 2 Tahapan analisis