EVALUASI KINERJA *SUPPORT VECTOR MACHINE* DAN *GRADIENT BOOSTING* DALAM ANALISIS SENTIMEN PUBLIK (STUDI KASUS: KEDATANGAN ETNIS ROHINGYA KE INDONESIA)

RIZQI NURSULISTIASARI

H011201022

PROGRAM STUDI MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2024

EVALUASI KINERJA *SUPPORT VECTOR MACHINE* DAN *GRADIENT BOOSTING* DALAM ANALISIS SENTIMEN PUBLIK (STUDI KASUS: KEDATANGAN ETNIS ROHINGYA KE INDONESIA)

RIZQI NURSULISTIASARI
H011201022

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024

EVALUASI KINERJA *SUPPORT VECTOR MACHINE* DAN *GRADIENT BOOSTING* DALAM ANALISIS SENTIMEN PUBLIK (STUDI KASUS: KEDATANGAN ETNIS ROHINGYA KE INDONESIA)

RIZQI NURSULISTIASARI

H011201022

Skripsi

sebagai salah satu syarat untuk memperoleh gelar sarjana

Program Studi Matematika

pada

PROGRAM STUDI MATEMATIKA
DEPARTEMEN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024

SKRIPSI

EVALUASI KINERJA *SUPPORT VECTOR MACHINE* DAN *GRADIENT BOOSTING* DALAM ANALISIS SENTIMEN PUBLIK (STUDI KASUS: KEDATANGAN ETNIS ROHINGYA KE INDONESIA)

RIZQI NURSULISTIASARI
H011201022

Skripsi,

telah dipertahankan di depan Panitia Ujian Sarjana Sains pada tanggal 6 Agustus 2024
dan dinyatakan telah memenuhi syarat kelulusan
pada

Program Studi Matematika
Departemen Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Hasanuddin
Makassar

Mengesahkan:
Pembimbing tugas akhir,

Dr. Khaeruddin, M.Sc.
NIP. 19650914 199103 1 003

Mengetahui:
Ketua Program Studi,

Dr. Firman, S.Si., M.Si.
NIP. 19680429 200212 1 001

PERNYATAAN KEASLIAN SKRIPSI DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa, skrpisi berjudul "Evaluasi Kinerja *Support Vector Machine* dan *Gradient Boosting* dalam Analisis Sentimen Publik (Studi Kasus: Kedatangan Etnis Rohingya ke Indonesia)" adalah benar karya saya dengan arahan dari bapak Dr. Khaeruddin, M.Sc. sebagai Pembimbing. Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka skripsi ini. Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini adalah karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut berdasarkan aturan yang berlaku.

Dengan ini saya melimpahkan hak cipta (hak ekonomis) dari karya tulis saya berupa skripsi ini kepada Universitas Hasanuddin.

Makassar, 6 Agustus 2024

Rizqi Nursulistiasari
H011201022

UCAPAN TERIMA KASIH

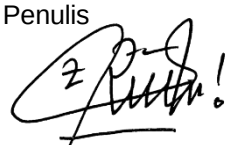
Alhamdulillah, segala puji syukur saya panjatkan kepada Allah SWT, yang dengan limpahan rahmat dan karunia-Nya, telah memberikan kekuatan dan ketabahan kepada saya untuk menyelesaikan skripsi ini. Sholawat serta salam semoga selalu tercurah kepada junjungan kita, Nabi Muhammad SAW, yang telah menjadi teladan hidup dan pembawa cahaya ilmu.

Dalam proses penyelesaian skripsi ini, saya dengan segala keterbatasan tentunya tidak dapat mencapai titik ini tanpa dukungan dan bantuan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, saya ingin mengucapkan terima kasih yang sebesar-besarnya kepada **kedua orang tua** saya. Tanpa do'a, motivasi, dan kasih sayang yang tulus, saya tidak akan mampu menghadapi setiap tantangan yang ada. Terima kasih atas kepercayaan dan dukungan yang tiada henti dalam setiap langkah dan keputusan yang saya ambil. Selain itu, teruntuk **adik dan keluarga** saya, terima kasih atas dukungan dan semangat yang tiada henti. Tidak lupa pula penulis sampaikan terima kasih yang mendalam kepada:

1. Bapak **Prof. Dr. Ir. Jamaluddin Jompa, M.Sc.** selaku Rektor Universitas Hasanuddin beserta jajarannya, Bapak **Dr. Eng. Amiruddin, M.Si.** selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, serta Bapak **Dr. Firman, S.Si., M.Si.** selaku Ketua Departemen Matematika. Terima kasih atas kebijakan yang telah memfasilitasi proses akademik saya selama ini.
2. Bapak **Dr. Khaeruddin, M.Sc.** selaku dosen pembimbing yang telah memberikan bimbingan, arahan, dan dukungan yang tiada henti selama proses penyusunan skripsi ini. Terima kasih atas waktu, ilmu, dan kesabaran yang telah Bapak berikan.
3. Bapak **Prof. Dr. Budi Nurwahyu, MS.** dan Bapak **A. Muh. Amil Siddik, S.Si., M.Si.** selaku dosen penguji, terima kasih atas waktu dan perhatian yang telah diluangkan untuk menilai dan memberikan masukan konstruktif bagi skripsi ini. Kritikan dan saran yang diberikan sangat membantu dalam meningkatkan kualitas penelitian ini.
4. Seluruh **Dosen dan Staf** Fakultas Matematika dan Ilmu Pengetahuan Alam, atas bantuan administratif dan dukungan yang telah diberikan selama proses akademik saya selama ini.
5. Sahabat-sahabat saya, **Sepia, Winda, Rahayu, Iin, Iyyu,** serta teman seperjuangan saya, **Ainun, Riska, dan Nisa** yang telah memberikan dukungan moral dan saling membantu dalam menghadapi berbagai tantangan.
6. Teman-teman **KKNT Pangkep Gelombang 110** khususnya **posko Bulu Cindea**, atas pengalaman berharga yang kita lalui bersama selama kegiatan KKN, yang turut memberikan inspirasi dan motivasi dalam penulisan skripsi ini.

Terakhir, ucapan terima kasih juga saya sampaikan kepada pihak-pihak lain yang tidak dapat saya sebutkan satu per satu. Semoga Allah SWT membalas segala kebaikan dan bantuan yang telah diberikan.

Penulis



Rizqi Nursulistiasari

ABSTRAK

RIZQI NURSULISTIASARI. **Evaluasi kinerja support vector machine dan gradient boosting dalam analisis sentimen publik (studi kasus: kedatangan etnis rohingya ke indonesia)** (dibimbing oleh Khaeruddin).

Latar belakang. Analisis sentimen telah menjadi topik yang sangat relevan dalam memahami pandangan dan tanggapan masyarakat terhadap berbagai peristiwa dan isu kontemporer. **Tujuan**. Penelitian ini bertujuan untuk mengevaluasi kinerja dua metode klasifikasi sentimen, yaitu *Support Vector Machine* (SVM) dan *Gradient Boosting* dalam menganalisis sentimen publik terkait kedatangan etnis Rohingya ke Indonesia berdasarkan data dari platform X (twitter). Data yang digunakan adalah *tweet* yang diposting pada tanggal 25 Desember 2023 hingga 28 Desember 2023 dengan kata kunci "Rohingya". **Metode**. Proses analisis meliputi *preprocessing data*, pelabelan sentimen menggunakan *library TextBlob*, pembagian data dengan metode *Hold-out Validation*, pembobotan kata menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF), penyeimbangan data dengan metode SMOTE, dan klasifikasi menggunakan *Support Vector Machine* dan *Gradient Boosting*. Evaluasi performa dilakukan dengan membandingkan akurasi, presisi, *recall*, dan *F1-score* dari kedua metode tersebut dengan menggunakan data uji. **Hasil**. Mayoritas *tweet* yang dianalisis mengandung sentimen netral, dengan *Gradient Boosting* menunjukkan kinerja yang lebih baik daripada *Support Vector Machine*. *Gradient Boosting* mencapai akurasi sebesar 91%, presisi 91%, *recall* 90%, dan *F1-score* 90%, sementara SVM memiliki akurasi sebesar 88%, presisi 88%, *recall* 87%, dan *F1-score* 87%. **Kesimpulan**. Temuan ini memberikan wawasan yang berharga mengenai pandangan publik terhadap kedatangan etnis Rohingya ke Indonesia dan menunjukkan potensi *Gradient Boosting* sebagai metode yang efektif dalam menganalisis sentimen publik.

Kata kunci: *Support Vector Machine*; *Gradient Boosting;* Akurasi; Presisi; *Recall*; *F1-score*

ABSTRACT

RIZQI NURSULISTIASARI. **Performance evaluation of support vector machine and gradient boosting in public sentiment analysis (case study: the arrival of ethnic rohingya to indonesia)** (supervised by Khaeruddin).

Background. Sentiment analysis has become a highly relevant topic in understanding people's views and responses to contemporary events and issues. **Aim**. This study aims to evaluate the performance of two sentiment classification methods, namely Support Vector Machine (SVM) and *Gradient Boosting* in analyzing public sentiment related to the arrival of Rohingya ethnicity to Indonesia based on data from platform X (twitter). The data used are tweets posted on December 25, 2023 to December 28, 2023 with the keyword “Rohingya”. **Methods**. The analysis process includes data preprocessing, sentiment labeling using TextBlob library, data sharing with Hold-out Validation method, word weighting using Term Frequency-Inverse Document Frequency (TF-IDF) method, data balancing with SMOTE method, and classification using Support Vector Machine and *Gradient Boosting*. Performance evaluation is done by comparing the accuracy, precision, *recall*, and *F1-score* of the two methods using test data. **Results**. The majority of tweets analyzed contained neutral sentiments, with *Gradient Boosting* showing better performance than Support Vector Machine. *Gradient Boosting* achieved 91% accuracy, 91% precision, 90% *recall*, and 90% *F1-score*, while SVM had 88% accuracy, 88% precision, 87% *recall*, and 87% *F1-score*. **Conclusion**. The findings provide valuable insights into public views on the Rohingya's arrival to Indonesia and demonstrate the potential of *Gradient Boosting* as an effective method in analyzing public sentiment.

Keywords: Support Vector Machine; *Gradient Boosting*; Accuracy; Precision; Recall; *F1-score*

DAFTAR ISI

Halaman

DAFTAR TABEL

Nomor Urut Halaman

DAFTAR GAMBAR

Nomor Urut Halaman

DAFTAR LAMPIRAN

Nomor Urut Halaman

BAB I

PENDAHULUAN

1.1 Latar Belakang

Krisis kemanusiaan yang melibatkan etnis Rohingya di Myanmar telah menimbulkan reaksi global dan dampak yang signifikan terhadap sejumlah negara penerima pengungsi (Khaled, 2021). Sebagai kelompok minoritas Muslim di Myanmar, etnis Rohingya telah menghadapi kekerasan dan diskriminasi yang sistematis di negara asal mereka. Konflik ini mencapai puncaknya pada serangkaian serangan militer pada tahun 2017 yang menyebabkan ribuan etnis Rohingya harus melarikan diri dan mencari perlindungan di negara-negara tetangga, termasuk Indonesia. Kondisi ini membentuk salah satu krisis kemanusiaan paling kejam dan berkepanjangan dalam sejarah dunia kontemporer (Ansar, 2020). Kedatangan etnis Rohingya ke Indonesia telah menjadi fokus perhatian dari berbagai segmen masyarakat. Hal ini menuai pro dan kontra di media sosial, sebagian masyarakat menunjukkan solidaritas dan empati terhadap penderitaan etnis Rohingya, dan sebagian lainnya menunjukkan ketidaksetujuan atau kekhawatiran terkait implikasi kedatangan pengungsi tersebut terhadap kehidupan sosial, ekonomi, dan keamanan di Indonesia.

Dalam era globalisasi dan kemajuan teknologi informasi, media sosial telah menjadi saluran utama di mana masyarakat menyuarakan pendapat dan berbagi informasi secara *real-time*, menjadi sumber data yang berharga untuk memahami dinamika opini publik terkait suatu masalah. Namun, dalam mengevaluasi volume komentar yang sangat besar, diperlukan analisis secara akurat untuk memilah sentimen positif dan negatif (Ardianto et al., 2020). Oleh karena itu, analisis sentimen atau *opinion mining* menjadi salah satu solusi untuk mengatasi masalah dalam mengklasifikasikan opini atau ulasan secara otomatis ke dalam kategori opini positif atau negatif (Indrayuni & Nurhadi, 2022). Analisis sentimen umumnya dijelaskan sebagai proses untuk menetapkan skor dan kategori sentimen, dengan merujuk pada kesesuaian kata kunci dan frasa dengan kamus skor sentimen serta leksikon yang telah disesuaikan (Samuel et al., 2020).

Penelitian terkait analisis sentimen di media sosial telah dilakukan pada berbagai isu penting dan kontroversial, mencerminkan keberagaman topik yang dapat dianalisis menggunakan teknik analisis sentimen. Misalnya, Vindua & Zailani (2023) meneliti sentimen terhadap Pemilu Indonesia 2024 yang dapat memberikan wawasan terkait pandangan dan perasaan publik terhadap proses pemilihan umum tersebut. Di sisi lain, Utomo et al., (2023) mengkaji respons publik terhadap Kebijakan Pemberlakuan Pembatasan Kegiatan Masyarakat (PPKM). Penelitian ini memberikan gambaran yang berharga tentang bagaimana kebijakan-kebijakan tersebut diterima dan dipahami oleh masyarakat luas. Selain isu-isu domestik, beberapa penelitian juga mengeksplorasi sentimen terkait konflik global yang mempengaruhi banyak pihak. Fajri et al. (2021) meneliti sentimen terhadap konflik Palestina-Israel, sementara Muriyatmoko et al. (2022) mengkaji respons publik terhadap konflik Rusia-Ukraina. Penelitian ini tidak hanya memberikan wawasan tentang pandangan dan perasaan publik terhadap isu-isu

internasional yang kompleks, tetapi juga menyoroti peran media sosial sebagai platform untuk ekspresi opini dan solidaritas.

Dalam konteks analisis sentimen, berbagai metode pembelajaran mesin (*machine learning*) telah dikembangkan dan digunakan untuk menganalisis data teks, diantaranya adalah *Support Vector Machine* (SVM) dan *Gradient Boosting*. Kedua metode ini telah terbukti efektif dalam mengklasifikasikan sentimen teks dengan tingkat akurasi yang tinggi. Penelitian yang dilakukan oleh Alzamzami et al. (2020) menunjukkan bahwa pengklasifikasi sentimen menggunakan *Light Gradient Boosting Machine* (LGBM) memiliki kinerja unggul dalam menangani data berdimensi tinggi dan tidak seimbang. Selain itu, Idris et al. (2023) menyimpulkan bahwa SVM terbukti sebagai algoritma klasifikasi yang cukup baik untuk menganalisis sentimen pada pengguna aplikasi shopee dengan perolehan akurasi sebesar 98% dan f1-score sebesar 98%.

Dari penelitian-penelitian tersebut, dapat disimpulkan bahwa baik SVM maupun *Gradient Boosting* memiliki keunggulan masing-masing dalam analisis sentimen. SVM dikenal karena kemampuannya dalam menghasilkan margin yang optimal antara kelas, yang membuatnya sangat efektif dalam mengklasifikasikan data yang memiliki pola yang kompleks dan bervariasi. Sementara *Gradient Boosting* unggul dalam membangun model yang kuat melalui penggabungan sejumlah model lemah serta memiliki fleksibilitas yang tinggi dalam menangani berbagai data yang tidak seimbang dan memiliki fitur non-linear yang kompleks. Pemilihan metode terbaik sangat tergantung pada karakteristik dataset dan kebutuhan spesifik dari analisis yang dilakukan. Meskipun kedua metode ini telah menunjukkan kinerja yang baik dalam berbagai aplikasi, belum banyak penelitian yang secara khusus mengevaluasi dan membandingkan kinerja SVM dan *Gradient Boosting* dalam konteks analisis sentimen publik terkait isu sensitif seperti kedatangan etnis Rohingya ke Indonesia.

Berdasarkan uraian sebelumnya, maka penelitian ini akan difokuskan untuk mengevaluasi kinerja *Support Vector Machine* dan *Gradient Boosting* dalam menganalisis sentimen publik terkait kedatangan etnis Rohingya ke Indonesia. Studi kasus ini diharapkan dapat memberikan wawasan yang lebih komprehensif tentang efektivitas kedua metode tersebut dalam menangani masalah analisis sentimen yang kompleks. Dengan membandingkan kinerja SVM dan *Gradient Boosting*, penelitian ini akan mengidentifikasi metode yang paling sesuai untuk digunakan dalam analisis sentimen di masa depan, khususnya dalam konteks isu-isu sosial yang sensitif.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dapat dirinci sebagai berikut:

1. Bagaimana sentimen pengguna media sosial X (twitter) terhapad kedatangan etnis Rohingya ke Indonesia?
2. Bagaimana perbandingan kinerja *Support Vector Machine* dan *Gradient Boosting* dalam mengklasisfikasikan sentimen pengguna X (twitter) terhadap kedatangan etnis Rohingya ke Indonesia?

1.3 Batasan Masalah

Untuk memastikan fokus penelitian yang jelas, beberapa batasan masalah diberlakukan:

1. Analisis sentimen hanya dilakukan terhadap data yang ditemukan pada platform X (twitter) dengan menggunakan kata kunci "Rohingya",
2. Penelitian ini hanya difokuskan pada ekspresi dan opini pengguna X (twitter) yang menggunakan bahasa Indonesia,
3. Data yang digunakan dalam penelitian ini terbatas pada unggahan *tweet* yang diposting dalam rentang waktu tanggal 25 Desember 2023 hingga 28 Desember 2023, dan
4. Metode analisis yang digunakan adalah *Support Vector Machine* dan *Gradient Boosting* yang nantinya akan dievaluasi berdasarkan akurasi, presisi, *recall*, dan *F1-score* dalam mengklasifikasikan sentimen.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk menganalisis dan memahami sentimen pengguna media sosial X (twitter) terkait kedatangan etnis Rohingya ke Indonesia, serta mengevaluasi dan membandingkan kinerja dua metode klasifikasi, yaitu *Support Vector Machine* (SVM) dan *Gradient Boosting* dalam mengklasifikasikan sentimen pengguna X (twitter) terhadap peristiwa tersebut.

1.5 Manfaat Penelitian

Manfaat penelitian ini dapat dibagi menjadi beberapa aspek yang mencakup berbagai pihak, diantaranya yaitu:

1. Manfaat Akademis bagi Mahasiswa:
   - Mahasiswa dapat memperoleh pemahaman yang mendalam tentang penggunaan metode klasifikasi teks, seperti *Support Vector Machine* dan *Gradient Boosting* dalam analisis sentimen, dan
   - Mahasiswa dapat mengembangkan keterampilan dalam melakukan penelitian ilmiah, pengumpulan data, analisis data, dan interpretasi hasil.
2. Manfaat Praktis bagi Peneliti dan Praktisi:
   - Peneliti dan praktisi di bidang analisis sentimen dapat menggunakan hasil penelitian ini sebagai referensi dalam memilih metode klasifikasi yang paling sesuai untuk analisis sentimen di berbagai konteks, dan
   - Penelitian ini dapat memberikan wawasan baru dan pemahaman yang lebih mendalam tentang efektivitas *Support Vector Machine* dan *Gradient Boosting* dalam menganalisis sentimen publik terkait isu-isu kontroversial.
3. Manfaat Sosial bagi Masyarakat:
   - Hasil penelitian ini dapat memberikan pemahaman yang lebih baik tentang pandangan dan opini masyarakat terhadap kedatangan etnis Rohingya ke Indonesia, dan
   - Informasi yang diperoleh dari penelitian ini dapat digunakan untuk merancang kebijakan atau langkah-langkah intervensi yang lebih tepat dan efektif dalam menangani isu-isu sosial yang sensitif.

Dengan demikian, penelitian ini diharapkan dapat memberikan manfaat yang signifikan baik dalam konteks akademis maupun praktis, serta memberikan kontribusi positif terhadap pemahaman dan penyelesaian masalah dalam masyarakat.

1.6 Landasan Teori

1.6.1 Etnis Rohingya

Etnis Rohingya adalah kelompok etnis minoritas yang berasal dari negara Myanmar, khususnya di wilayah Rakhine. Sejarah panjang mereka di Myanmar ditandai oleh konflik etnis dan agama yang kompleks, di mana mereka telah menghadapi diskriminasi sistemik dan pembatasan hak-hak dasar. Operasi militer yang menargetkan etnis Rohingya telah berlangsung secara berulang, khususnya pada tahun 1978, 1991-1992, dan yang paling baru terjadi pada tahun 2017-2018 (Rahman et al., 2020). Serangan ini menyebabkan ribuan warga Rohingya tewas, desa-desa mereka dihancurkan, dan telah menciptakan kondisi yang tidak layak untuk hidup, sehingga mendorong banyak orang Rohingya untuk mencari perlindungan di luar negeri.

Status dan hak warga Rohingya terus menjadi perhatian utama dalam diplomasi internasional dan upaya kemanusiaan global. Pemerintah Myanmar menolak mengakui kewarganegaraan etnis Rohingya dengan alasan bahwa mereka bukan bagian dari kelompok etnis yang sudah ada di Myanmar sebelum kemerdekaan pada tahun 1948 (Ruslan et al., 2023). Undang-undang tahun 1982 tentang kewarganegaraan mengecualikan mereka sebagai warga negara nasional, dan sepanjang sejarah, mereka berulang kali dipindahkan ke negara tetangga Bangladesh oleh negara (D'Silva, 2021).

1.6.2 X (Twitter)

Twitter atau yang sekarang telah berganti nama dengan sebutan X adalah salah satu platform media sosial yang dioperasikan oleh perusahaan X Corp. sebagai penerus Twitter, Inc. Twitter didirikan pada Juli 2006 oleh Jack Dorsey, Noah Glass, Bizz Ston, dan Evan Williams, lalu resmi berganti nama menjadi X pada Juli 2023. Twitter adalah platform media sosial populer yang memungkinkan pengguna untuk menyuarakan pendapat mereka terkait peristiwa global secara *real-time* melalui pesan singkat yang disebut *tweet* (Sheikh & Jaiswal, 2020). Pentingnya twitter sebagai alat untuk menyuarakan pendapat juga terletak pada kecepatan penyebaran informasi. Berita atau pendapat yang diunggah dapat dengan cepat menjadi topik hangat dan mendapat perhatian luas di seluruh dunia. Oleh karena itu, basis data twitter telah menjadi area penelitian yang sangat menarik bagi para peneliti (Siddiqui et al., 2021). Namun, perlu diingat bahwa twitter juga memiliki tantangan seperti disinformasi, kebencian daring, dan polemik yang dapat muncul sebagai akibat dari kebebasan berpendapat yang luas.

1.6.3 Analisis Sentimen

Sentimen adalah sikap, pandangan, atau perasaan seseorang terhadap suatu objek, peristiwa, atau subjek tertentu yang dapat bersifat positif, negatif, atau netral. Dalam konteks analisis teks, sentimen diekspresikan melalui kata-kata yang digunakan dalam tulisan, seperti dalam ulasan produk, komentar di media sosial, artikel berita, dan lain-lain. Sedangkan analisis sentimen adalah bidang studi yang menganalisis pendapat, sentimen, penilaian, evaluasi sikap, dan emosi individu terhadap suatu topik, layanan, produk, individu, organisasi, atau aktivitas khusus (Hamidi et al., 2021). Analisis sentimen merupakan komponen dari analisis teks yang berada dalam cakupan pemrosesan bahasa alami (NLP). Tujuan utama analisis sentimen adalah memahami polaritas

pendapat yang diungkapkan dalam ulasan dan jenis emosi terhadap berbagai aspek subjek, baik positif maupun negatif (Bahaaaldeen Abdul wahhab & KareemabdulHassan, 2019). Metode analisis sentimen dapat dikategorikan menjadi dua pendekatan utama yaitu, metode berbasis kamus dan metode berdasarkan pembelajaran mesin (Liu et al., 2022).

1. Metode Berbasis Kamus (*Lexicon-Based Methods*)
   Metode berbasis kamus menggunakan daftar kata atau kamus yang sudah dilabeli dengan sentimen positif, negatif, atau netral. Proses analisis dilakukan dengan mencocokkan kata-kata dalam teks dengan entri kamus dan mengkalkulasi skor sentimen berdasarkan frekuensi dan bobot kata-kata tersebut. Pendekatan ini tidak memerlukan pelatihan model dan cenderung lebih sederhana, tetapi bisa kurang fleksibel dan tidak selalu akurat dalam menangkap konteks atau nuansa yang lebih kompleks. Adapun contoh kamus sentimen yang dapat digunakan adalah SentiWordNet dan AFINN
2. Metode Berdasarkan Pembelajaran Mesin (*Machine Learning-Based Methods*)
   Metode berbasis pembelajaran mesin melibatkan pelatihan model menggunakan data yang telah diberi label untuk mengklasifikasikan sentimen. Metode ini dapat menggunakan teknik-teknik dari pembelajaran mesin tradisional maupun deep learning untuk menganalisis dan memprediksi sentimen. Proses ini biasanya mencakup beberapa tahapan seperti ekstraksi fitur, pelatihan model, dan evaluasi performa. Adapun contoh teknik yang dapat digunakan adalah *Support Vector Machine* (SVM) dan *Gradient Boosting*.

Selain itu, metode analisis sentimen dapat diklasifikasikan lebih lanjut menjadi analisis sentimen eksplisit dan implisit, dengan analisis sentimen eksplisit yang berfokus pada sentimen yang diungkapkan langsung dalam teks, dan analisis sentimen implisit yang menyimpulkan sentimen dari konteks atau indikator tidak langsung lainnya (M. Chen et al., 2022).

1.6.4 *Text Mining*

Text mining adalah proses pengambilan data, informasi, dan pengetahuan dari data tekstual yang tidak terstruktur dengan mengubahnya ke dalam format yang dapat dimengetri oleh mesin dan menyajikan hasilnya dalam bentuk yang mudah digunakan (Isaeva & Aldarova, 2021). *Text mining* dapat didefinisikan secara luas sebagai suatu metode eksplorasi informasi dimana pengguna berinteraksi dengan sejumlah dokumen menggunakan tools analisis dari komponen-komponen dalam *data mining* yang salah satunya adalah kategorisasi (Alhaq et al., 2021). Teknik ini menggunakan metode dari *machine learning*, statistik, dan linguistik untuk mengubah teks menjadi data yang dapat dianalisis. Tujuan *Text mining* adalah mengubah teks menjadi data terstruktur yang dapat dianalisis untuk mendapatkan pemahaman dan informasi yang bermanfaat dengan cara mengidentifikasi pola linguistik yang khas dari sumber data (Irawan et al., 2022). Selain itu, *text mining* juga dapat digunakan untuk mengelompokkan teks ke dalam kategori atau topik tertentu, memungkinkan pemahaman yang lebih baik tentang tren atau isu tertentu dalam teks besar.

1.6.5 *Machine Learning*

Machine learning adalah cabang dari kecerdasan buatan (*Artificial Intelligence*) yang berkembang pesat dan telah menjadi topik utama dalam dunia teknologi. Teknik-teknik *Machine learning* memungkinkan komputer untuk belajar dari data yang ada dan membuat keputusan atau melakukan prediksi berdasarkan pola-pola yang terdapat dalam data tanpa harus secara eksplisit diprogram oleh manusia (Wang et al., 2023). Salah satu konsep kunci dalam *Machine learning* adalah penggunaan algoritma untuk melatih model-model yang dapat mengenali pola-pola tersebut dan membuat prediksi yang akurat (Hagmann & Riezler, 2021). Dalam konteks analisis sentimen publik, *machine learning* memainkan peran penting dalam mengolah data teks dari media sosial, berita, dan sumber-sumber lain untuk memahami opini dan perasaan masyarakat terhadap isu-isu tertentu, seperti kedatangan etnis Rohingya ke Indonesia. *Machine learning* dapat dibagi menjadi beberapa jenis berdasarkan cara pembelajaran model dari data (Shweta Pandey et al., 2023):

1. *Supervised learning*
   Dalam *supervised learning*, model dilatih menggunakan dataset yang telah dilabeli, yaitu setiap contoh input terkait dengan output yang benar. Tujuan model adalah mempelajari hubungan antara input dan output sehingga dapat memprediksi output yang benar untuk input baru.
2. *Unsupervised learning*
   Dalam *unsupervised learning*, model dilatih menggunakan data yang tidak dilabeli. Tujuan utama adalah menemukan struktur atau pola tersembunyi dalam data. Algoritma yang sering digunakan termasuk clustering dan reduksi dimensi.
3. *Reinforcement learning*
   Dalam *reinforcement learning*, model belajar melalui interaksi dengan lingkungan dan menerima umpan balik dalam bentuk *reward* atau *punishment*. Model membuat keputusan berurutan dengan tujuan memaksimalkan reward kumulatif.

Dalam konteks analisis sentimen publik, *supervised learning* adalah jenis yang paling relevan karena model dilatih menggunakan dataset yang telah dilabeli, di mana setiap contoh teks terkait dengan label sentimen (positif, negatif, atau netral).

1.6.6 *Data crawling*

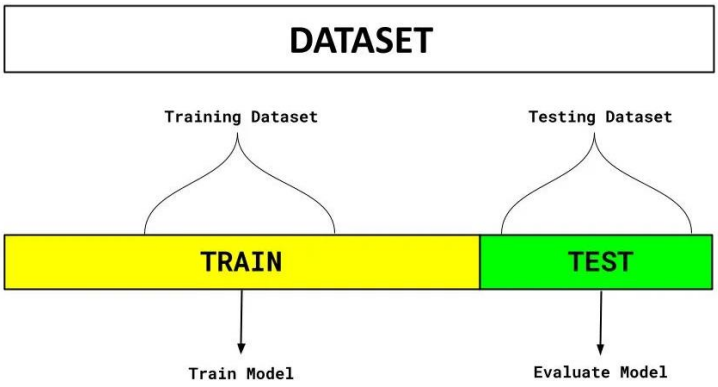
Data crawling adalah suatu teknik yang secara otomatis mengumpulkan data dari suatu situs web berdasarkan kata kunci yang dimasukkan oleh pengguna (Dikiyanti et al., 2021). Tujuan utama dari *data crawling* adalah untuk mengindeks dan mengumpulkan data yang dapat digunakan dalam berbagai tujuan, seperti analisis, penelitian, atau pembangunan basis data. Proses ini memungkinkan pengumpulan informasi secara otomatis dari internet dengan cepat dan efisien, memungkinkan organisasi atau peneliti untuk memanfaatkan data besar yang tersebar di berbagai situs web (Rismawan & Syahidin, 2023). *Data crawling* memiliki aplikasi luas dalam dunia digital, seperti dalam pengumpulan data untuk mesin pencari, pengawasan harga *e-commerce*, atau pengumpulan berita dan informasi terkini. Meskipun memiliki manfaat yang besar, *data crawling* juga memerlukan manajemen dan pemantauan yang cermat agar tidak mengganggu kinerja server situs web yang dijelajahi atau melanggar kebijakan penggunaan data.

1.6.7 *Text Preprocessing*

Text Preprocessing adalah bagian penting dari *text mining* yang melibatkan pembersihan dan persiapan data teks sebelum analisis lebih lanjut. *Text Preprocessing* merupakan tahap krusial dalam klasifikasi teks yang dapat membantu meningkatkan kualitas data dan memastikan bahwa teks dapat diolah secara efektif oleh model NLP atau algoritma lainnya (Shruthi & Anil Kumar, 2020). *Preprocessing* bertujuan untuk membersihkan, merapihkan, dan mempersiapkan teks mentah agar dapat diolah secara efektif oleh algoritma komputer. Proses ini melibatkan beberapa tahapan, di antaranya *case folding*, *cleaning*, *normalization*, *tokenization*, *stopword removal*, dan *stemming* (Hashfi et al., 2022).

1.6.8 *Hold-Out Validation*

Hold-out validation merupakan salah satu metode validasi yang umum digunakan dalam pembangunan model *machine learning*. Metode ini membagi dataset menjadi dua bagian, yaitu data latih (*training set*) dan data uji (*test set*), dengan proporsi tertentu (Tempola et al., 2021). Proporsi pembagian data antara data latih dan data uji dapat bervariasi tergantung pada ukuran dataset dan kompleksitas model yang dibangun. Secara umum, proporsi yang umum digunakan adalah 70-80% data latih dan 20-30% data uji. Proporsi ini memastikan bahwa model dilatih dengan cukup data untuk belajar secara efektif, sementara masih memiliki data yang cukup untuk menguji kinerjanya.

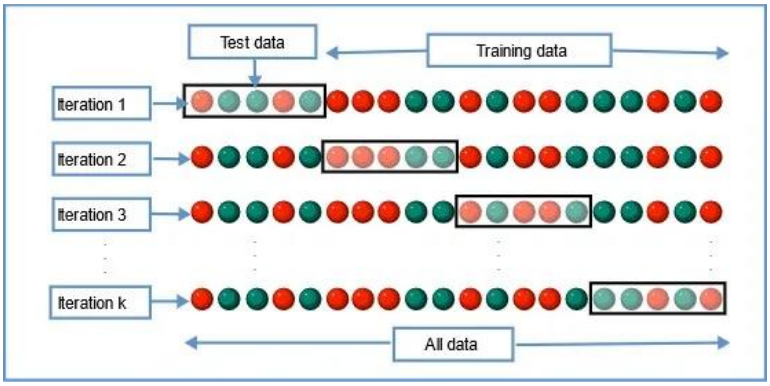


Gambar 1. 1 Pembagian data dengan *Hold-Out Validation*

Metode *hold-out validation* memiliki beberapa kelebihan, termasuk kemudahan implementasi dan interpretasi. Pembagian dataset menjadi dua bagian yang saling eksklusif membuat evaluasi kinerja model menjadi lebih sederhana. Selain itu, *hold-out validation* cocok digunakan untuk dataset yang besar dan kompleks (Santoso & Wibowo, 2022). Meskipun memiliki kelebihan, *hold-out validation* juga memiliki beberapa kekurangan. Proporsi pembagian data dapat mempengaruhi hasil evaluasi model, terutama jika dataset memiliki distribusi kelas yang tidak seimbang. Selain itu, performa model dapat bervariasi tergantung pada bagaimana data dibagi, sehingga hasil evaluasi mungkin tidak konsisten.

1.6.9 *Cross Validation*

Cross Validation adalah teknik evaluasi model yang digunakan secara luas dalam pembelajaran mesin untuk mengukur kinerja model dan memastikan bahwa model tersebut dapat digeneralisasi dengan baik terhadap data yang tidak terlihat. Teknik ini membantu dalam mengurangi *overfitting* dan *underfitting* (You et al., 2023), serta memberikan estimasi yang lebih akurat tentang kinerja model di masa mendatang (Purba et al., 2022). *Cross Validation* sangat penting dalam proses pemodelan karena membantu dalam memilih model yang paling sesuai dan menyesuaikan *hyperparameter* dengan cara yang lebih andal dibandingkan dengan metode evaluasi tradisional. *Cross Validation* melibatkan pembagian dataset asli menjadi beberapa subset atau lipatan. Setiap lipatan digunakan bergantian sebagai set pengujian, sementara lipatan yang tersisa digunakan sebagai set pelatihan (Nur-a-alam et al., 2021). Proses ini diulang beberapa kali sehingga setiap lipatan digunakan sebagai set pengujian satu kali. Hasil kinerja dari setiap iterasi kemudian dirata-rata untuk memberikan estimasi kinerja keseluruhan model.



Gambar 1. 2 Pembagian data dengan *Cross Validation*

Cross Validation memiliki berbagai variasi diantaranya yaitu:

1. *K-Fold Cross Validation*

   K-Fold Cross Validation memastikan bahwa setiap observasi dalam dataset digunakan untuk pengujian dan pelatihan, menghasilkan estimasi kinerja model yang lebih stabil dan akurat. Dataset dibagi menjadi k subset seukuran yang sama. Setiap subset digunakan satu kali sebagai data uji, sementara k-1 subset lainnya digunakan sebagai data latih. Berikut adalah rumus yang digunakan.

$$MSE_{CV} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i \tag{1.1}$$

   Keterangan:

   $\mathrm{MSE}_{\mathrm{CV}}$ = *Mean Squared Error* dari model yang dihitung,

   k = Jumlah lipatan (*folds*),

   MSE_i = *Mean Squared Error* pada *fold* ke-*i.*

   Nilai MSE_i pada pesamaan 1.1 dapat diketahui menggunakan rumus berikut:

$$\mathrm{MSE}_i = \frac{1}{n_i}\sum_{j \in Fold_i} \left(y_j - \hat{y}_j\right)^2$$

Keterangan:

MSE_i = *Mean Squared Error* pada *fold* ke-*I,*

n_i = Jumlah sampel dalam lipatan ke-*i.*

$(y_j - \hat{y}_j)^2$ = Kuadrat dari selisih antara nilai sebenarnya y_j dan nilai prediksi $\hat{y}_j$ untuk sampel *j* dalam lipatan ke-*i*.

2. *Stratified K-Fold Cross Validation*

   Stratified K-Fold Cross Validation adalah variasi dari *K-Fold Cross Validation* yang memastikan proporsi kelas yang seimbang di setiap *fold*. *Stratified K-Fold Cross Validation* penting untuk dataset dengan distribusi kelas yang tidak merata, memastikan evaluasi model yang lebih objektif dan akurat. Berikut adalah rumus yang digunakan:

$$\text{Stratified } MSE_{CV} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

   Keterangan:

   $\text{Stratified } \mathrm{MSE}_{\mathrm{CV}}$ = Rata-rata dari *Mean Squared Error* di seluruh lipatan (*folds*),

   k = Jumlah lipatan (*folds*),

   MSE_i = *Mean Squared Error* pada *fold* ke-*i.*

3. *Time Series Cross Validation*

   Time Series Cross Validation penting untuk memodelkan perilaku data dalam urutan waktu, memastikan model dapat diprediksi dan relevan untuk penggunaan di masa depan. Data dari masa lalu digunakan untuk pelatihan dan data yang lebih baru digunakan untuk pengujian. Tidak ada rumus matematis khusus, tetapi fokus pada pengujian model dengan data yang terus bergerak maju dalam waktu.

4. *Leave-One-Out Cross Validation* (LOOCV)

   LOOCV memberikan estimasi kinerja model yang paling akurat karena menggunakan hampir semua data untuk pelatihan dan evaluasi. Setiap observasi secara berurutan dijadikan data uji, sementara sisanya digunakan sebagai data latih. Tidak ada rumus matematis khusus karena LOOCV menghitung metrik evaluasi untuk setiap observasi secara individu.

Pemilihan jenis *Cross Validation* yang sesuai harus mempertimbangkan karakteristik dataset, termasuk distribusi kelas, struktur waktu, dan ukuran dataset. Meskipun memiliki beberapa keterbatasan, manfaat yang diberikan oleh *Cross Validation* dalam hal pengurangan *overfitting*, penggunaan data yang efisien, dan estimasi kinerja yang lebih baik menjadikannya alat yang sangat berharga dalam *toolkit* pembelajaran mesin.

1.6.10 *Term Frequency – Inverse Document Frequency (TF-IDF)*

Term Frequency – Inverse Document Frequency (TF-IDF) adalah metode statistik numerik yang digunakan untuk mengukur dan menilai relevansi suatu kata dalam suatu dokumen terhadap seluruh koleksi dokumen dengan memberikan bobot untuk setiap kata (Wulandari et al., 2021). TF-IDF digunakan untuk mewakili dokumen dalam bentuk vektor, di mana setiap dimensi vektor mencerminkan bobot relatif kata-kata. Hal ini memungkinkan pemodelan data teks dalam bentuk numerik yang dapat digunakan dalam berbagai algoritma *Machine learning*. Metode ini mengintegrasikan dua konsep

dalam perhitungan bobot, yakni frekuensi kemunculan kata dalam suatu dokumen yang disebut dengan *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) untuk mengukur kemunculan suatu kata pada keseluruhan dokumen. (Sari et al., 2021).

1. *Term Frequency* (TF)

$$tf(t, D_j) = \frac{f_{t,d_j}}{\sum_i f_{i,d_j}} \quad (1.2)$$

Keterangan:

t = *Term* atau kata yang diukur frekuensinya,
D_j = Dokumen ke-j dalam koleksi dokumen,
f_{t,d_j} = Frekuensi kemunculan term *t* dalam dokumen *d*,
$\sum_i f_{i,d}$ = Total jumlah term dalam dokumen *d*.

2. *Inverse Document Frequency* (IDF)

$$idf(t, D_j) = \log\left(\frac{N+1}{df_t+1}\right) + 1 \quad (1.3)$$

Keterangan:

N = Total dokumen dalam koleksi,
df_t = Jumlah dokumen yang mengandung kata t dalam koleksi dokumen.

3. TF-IDF

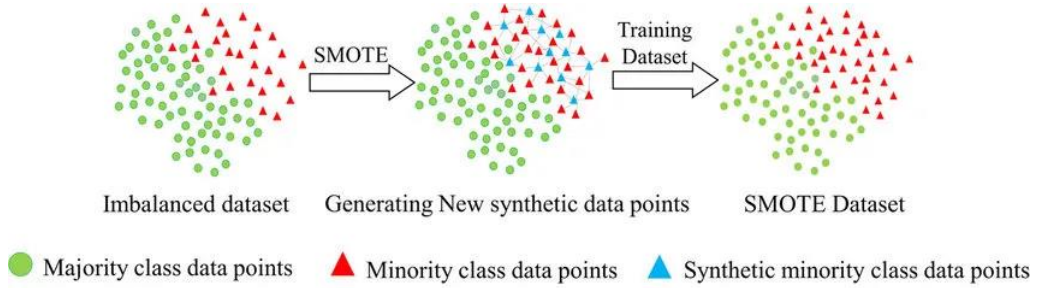
$$TF - IDF(t, d, D) = tf(t, d) \times idf(t, D)$$

Penerapan TF-IDF membantu menyoroti kata-kata kunci yang mungkin memiliki nilai informatif tinggi dalam suatu dokumen dan kumpulan dokumen secara keseluruhan. Pendekatan ini memberikan representasi vektor berbobot untuk kata-kata dalam dokumen yang dapat digunakan untuk analisis lebih lanjut, seperti pengelompokan atau klasifikasi dokumen. Keuntungan dari penggunaan TF-IDF adalah dapat memberikan bobot lebih tinggi kepada kata-kata yang memberikan informasi yang lebih spesifik tentang suatu dokumen daripada kata-kata umum yang mungkin muncul di seluruh kumpulan dokumen. Hal ini membantu dalam mengekstrak dan menyoroti kata-kata kunci atau istilah yang khas untuk dokumen tertentu dalam analisis teks atau pengambilan informasi.

1.6.11 SMOTE

SMOTE (*Synthetic Minority Oversampling Technique*) adalah metode *oversampling* yang dikembangkan untuk menangani ketidakseimbangan kelas dalam kumpulan data. SMOTE dikembangkan oleh Chawla dan telah menjadi salah satu teknik yang paling banyak digunakan untuk menangani masalah ketidakseimbangan kelas data (Rachmatullah, 2022). Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam satu kelas jauh lebih besar daripada jumlah sampel dalam kelas lainnya. Hal ini sering menyebabkan algoritma pembelajaran mesin memiliki bias terhadap kelas mayoritas dan mengabaikan kelas minoritas. Beberapa metode untuk menangani ketidakseimbangan kelas antara lain adalah *undersampling* kelas mayoritas, *oversampling* kelas minoritas, dan penggunaan algoritma yang dirancang khusus untuk menangani ketidakseimbangan.

Tidak seperti metode *oversampling* sederhana yang hanya menduplikasi sampel minoritas, SMOTE memperluas dataset dengan menghasilkan sampel sintetik berdasarkan sampel yang ada, sehingga meningkatkan representasi kelas minoritas dalam dataset (Chang et al., 2021). Penelitian komparatif menunjukkan bahwa SMOTE lebih unggul dibandingkan dengan metode *oversampling* lainnya, menghasilkan peningkatan yang signifikan dalam hal akurasi dan prediksi (Taghizadeh-Mehrjardi et al., 2020).



Gambar 1. 3 Cara kerja SMOTE

Berikut tahapan serta rumus yang digunakan dalam proses penerapan SMOTE:

1. Identifikasi kelas minoritas yang memiliki jumlah sampel lebih sedikit dibandingkan dengan kelas mayoritas.
2. Pemilihan TetanggaTerdekat (*k-Nearest Neighbors*)
   Untuk setiap sampel dari kelas minoritas, kita menemukan k sampel terdekat (*nearest neighbors*) dari kelas yang sama. Jarak antara sampel-sampel ini dihitung menggunakan metrik seperti *Euclidean distance*. Untuk setiap sampel minoritas x_i, jarak d ke setiap sampel minoritas lainnya dihitung. Selain itu, *k* tetangga terdekat dipilih berdasarkan jarak. Ini membantu dalam menentukan bagaimana sampel sintetis akan dibangun.

   $$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

   Keterangan:
   (x, y) = Vektor dengan n dimensi,
   $(x_i - y_i)^2$ = Kuadrat dari selisih komponen ke-*i* antara vektor x dan y.
3. Pembuatan Sampel Sintetis
   Untuk setiap sampel minoritas x_i, sampel sintetis baru x_{syn} dibuat dengan cara menggabungkan x_i dengan salah satu dari k tetangga terdekat secara acak. Proses ini dilakukan dengan menggunakan rumus berikut:

   $$x_{syn} = x_i + \delta \cdot (x_{knn} - x_i)$$

   Keterangan:
   x_i = Vektor asli atau referensi,
   x_{knn} = Vektor tetangga terdekat (*nearest neighbor*) dari x_i,
   δ = Faktor skalar yang mengontrol seberapa jauh x_{syn} dari x_i menuju x_{knn},

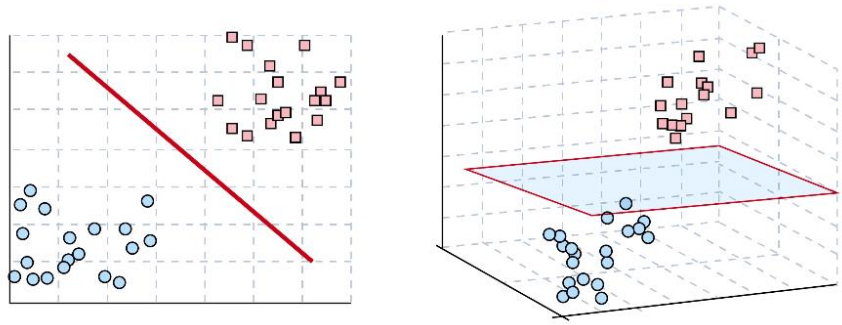
x_{syn} = hasil interpolasi linier antara x_i dan x_{knn}, di mana posisi x_i dipindahkan sejauh δ menuju x_{knn}.

4. Penambahan sampel sintetis yang dihasilkan ke dalam dataset asli, sehingga meningkatkan jumlah sampel kelas minoritas.

Seiring dengan popularitas SMOTE, berbagai variasi dan perbaikan telah diusulkan untuk meningkatkan kinerjanya, seperti *Borderline*-SMOTE, ADASYN (*Adaptive Synthetic Sampling*), dan SMOTE-NC (SMOTE untuk data kategori). Dalam praktiknya, SMOTE telah terbukti berguna dalam berbagai aplikasi, menjadi pilihan utama bagi para peneliti dan praktisi dalam menangani masalah ketidakseimbangan kelas.

1.6.12 *Support Vector Machine*

Support Vector Machine (SVM) adalah salah satu algoritma pembelajaran mesin yang populer dan efektif dalam klasifikasi dan regresi. SVM dikenal karena kemampuannya dalam menangani pemisahan kelas yang kompleks dan data yang tidak linear. Algoritma ini pertama kali diperkenalkan kepada komunitas ilmu komputer pada tahun 1990-an oleh Vapnik (1995) dan telah menjadi salah satu algoritma yang paling banyak digunakan dalam masalah klasifikasi (Montesinos López et al., 2022). Prinsip dasar SVM adalah membangun sebuah garis atau bidang yang memisahkan dua kelompok data dengan jarak terbesar di antara mereka, yang disebut margin. (Y. Chen & Yang, 2021). Garis atau bidang pemisah ini disebut *hyperplane.* Tujuannya adalah untuk memaksimalkan margin, yaitu jarak antara *hyperplane* dengan titik data terdekat dari setiap kelas, yang disebut *support vectors*. Pada kasus data yang tidak bisa dipisahkan secara linear, SVM menggunakan teknik yang disebut *kernel trick*. Teknik ini mengubah data ke dalam dimensi yang lebih tinggi sehingga data tersebut dapat dipisahkan dengan garis lurus atau bidang.



Gambar 1. 4 Cara kerja *Support Vector Machine*

Berikut tahapan dan rumus yang digunakan dalam SVM:

1. *Hyperplane* dan Margin
   SVM bekerja dengan mencari *hyperplane* yang memisahkan dua kelas data dengan margin terbesar. *Hyperplane* adalah sebuah garis atau bidang yang berfungsi sebagai pemisah antara dua kelas. Untuk data dua dimensi, *hyperplane* dapat diwakili oleh persamaan:

$$w \cdot x + b = 0$$

Keterangan:

w = Vektor bobot yang menentukan arah dan orientasi dari *hyperplane*,

x = Vektor fitur yang mewakili data input yang ingin diklasifikasikan,

b = Nilai bias atau offset yang menentukan seberapa jauh *hyperplane* dari titik asal.

2. Margin

Margin adalah jarak geometris antara *hyperplane* dengan titik data terdekat (support vectors) dari masing-masing kelas. Semakin besar margin, maka semakin baik model SVM dalam memisahkan data. Margin dapat didefinisikan sebagai:

$$\gamma = \frac{2}{\|w\|}$$

3. Fungsi Tujuan

Fungsi Tujuan adalah fungsi matematis yang didefinisikan untuk optimisasi dalam SVM. Tujuannya adalah untuk menemukan vektor bobot w dan bias b yang menghasilkan *hyperplane* dengan margin maksimal, yang direpresentasikan dengan meminimalkan norma.

- *Hard Margin*

  Hard margin SVM digunakan ketika data benar-benar dapat dipisahkan secara linear tanpa kesalahan. Ini berarti tidak ada titik data yang melanggar batas margin, dan semua titik data berada di sisi yang benar dari *hyperplane*. Untuk mencari *hyperplane* optimal, SVM meminimalkan fungsi tujuan berikut:

  $$\min_{w,b} \frac{1}{2}\|w\|^2$$

  Dengan kendala $y_i(w \cdot x_i + b) \geq 1, \ \forall i \in \{1, \ldots, n\}$

  di mana y_i adalah label kelas untuk sampel ke-i (1 atau -1), dan x_i adalah vektor fitur untuk sampel ke-i.

- *Soft Margin*

  Soft margin SVM digunakan ketika data tidak sepenuhnya dapat dipisahkan secara linear. Dalam kasus ini, kita memperkenalkan variabel slack ξ_i yang memungkinkan beberapa titik data berada di sisi yang salah dari margin atau *hyperplane*. Soft margin SVM berusaha untuk menemukan keseimbangan antara memaksimalkan margin dan meminimalkan kesalahan klasifikasi. Fungsi tujuan menjadi:

  $$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

  Dengan kendala $y_i(w \cdot x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \ \forall i \in \{1, \ldots, n\}$

  Di mana ξ_i adalah slack variables yang memperbolehkan *misclassifications*, dan C adalah *hyperparameter* regulasi yang mengontrol *trade-off* antara memaksimalkan margin dan meminimalkan kesalahan klasifikasi.

4. *Kernel Trick*

SVM dapat diubah untuk menangani data yang tidak linear dengan menggunakan kernel trick. Kernel adalah fungsi yang mengubah data ke dalam ruang dimensi tinggi, sehingga data dapat dipisahkan dengan *hyperplane* dalam ruang tersebut. Beberapa kernel yang umum digunakan adalah:

- Linear

$$K(x_i, x_j) = x_i \cdot x_j$$

- Polynomial

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

- Radial Basis Function (RBF)

$$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right)$$

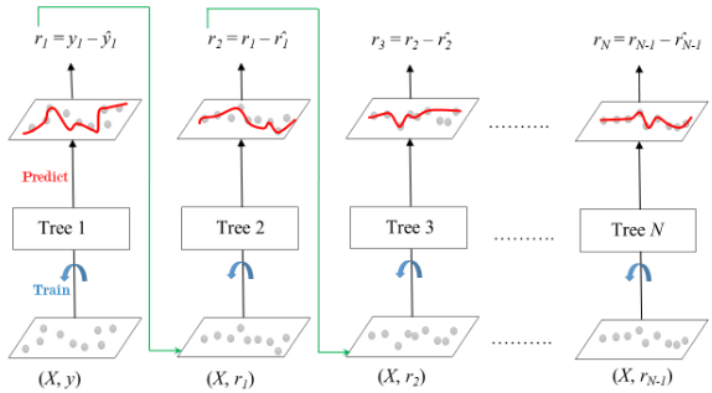
- Sigmoid

$$K(x_i, x_j) = \tanh(\alpha x_i \cdot x_j + r)$$

1.6.13 *Gradient Boosting*

Gradient Boosting merupakan teknik pembelajaran mesin yang digunakan untuk membuat model yang kuat dengan menggabungkan beberapa model yang lebih lemah. Algoritma ini secara khusus mengandalkan pohon keputusan sebagai model dasar untuk membentuk pengklasifikasi ansambel yang kuat, sehingga sering kali disebut sebagai pohon keputusan yang ditingkatkan gradien-nya (*Gradient Boosted Decision Tree*, GBDT). Teknik *Gradient Boosting* pertama kali diperkenalkan oleh Leo Breiman, yang mencatat bahwa *boosting* dapat direpresentasikan sebagai teknik optimasi pada fungsi kerugian yang sesuai. Selanjutnya, Jerome Friedman mengembangkan teknik ini lebih lanjut, yang mengarah pada penciptaan algoritma seperti *Gradient Boosting Machine* (Santoso & Wibowo, 2022). Lebih baru, teknik *Gradient Boosting* telah diimplementasikan dalam beberapa *algoritma* terkenal seperti *XGBoost* (*Extreme Gradient Boosting*), *LightGBM* (*Light Gradient Boosting Machine*), dan *CatBoost* (Ali et al., 2022).

Gradient Boosting bekerja dengan membangun model secara bertahap, di mana setiap model baru berusaha memperbaiki kesalahan yang dibuat oleh model sebelumnya. Setiap model baru dioptimalkan untuk mengurangi kesalahan menggunakan fungsi *loss*, yaitu suatu cara untuk mengukur seberapa jauh prediksi model dari nilai sebenarnya. Proses ini menggunakan metode *gradient descent*, yang membantu menemukan cara terbaik untuk memperbaiki kesalahan. Pada setiap iterasi, model baru dibuat untuk memperkirakan kesalahan (*residual*) dari prediksi model sebelumnya. *Residual* ini adalah selisih antara prediksi model dan nilai sebenarnya. Hasil prediksi dari model baru ini kemudian ditambahkan ke model yang sudah ada, sehingga model keseluruhan menjadi lebih akurat dengan mengurangi kesalahan keseluruhan.



Gambar 1. 5 Cara kerja *Gradient Boosting*

Secara umum, tahap-tahap dari algoritma *Gradient Boosting* adalah sebagai berikut:

1. Inisialisasi prediksi awal ($F_0(x)$)
   Menemukan prediksi awal ($F_0(x)$) yang paling sederhana atau model yang memberikan kesalahan paling kecil pada dataset pelatihan.
   $$F_0(x) = arg\ \min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$$
   Keterangan:
   $F_0(x)$: model awal atau model ansambel pada iterasi pertama ($m = 0$).
   N : Jumlah total sampel dalam dataset pelatihan
   $L(y_i, \gamma)$: Fungsi kerugian (*loss function*) antara label sebenarnya (y_i) dan prediksi awal (γ)
2. Untuk iterasi $m = 1, 2, 3, \ldots, M$
   a. Hitung residual (*pseudo-residuals*)
      Mengevaluasi seberapa besar kesalahan yang dibuat oleh model ansambel saat ini.
      $$r_i^{(m)} = -\left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right]$$
      Keterangan:
      $r_i^{(m)}$: Residual atau kesalahan pada sampel ke-i pada iterasi ke-m
      L : Fungsi kerugian yang digunakan untuk mengukur kesalahan prediksi.
      y_i : Label sebenarnya dari sampel ke-i.
      $F_{m-1}(x_i)$: Prediksi model pada iterasi sebelumnya terhadap sampel ke-i.
   b. Cocokkan model tambahan $h_m(x)$ ke *pseudo-residuals*
      Menemukan model tambahan ($h_m(x)$) yang dapat mengoreksi kesalahan dari model ansambel saat ini ($F_{m-1}(x_i)$). Model tambahan ini dapat berupa model sederhana seperti pohon keputusan yang ditentukan secara heuristik.
      $$h_m(x)\ = arg\ \min_{h} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \gamma h(x_i))$$
      Keterangan:
      $h_m(x)$: Model tambahan yang akan dipilih pada iterasi ke-m
      $\gamma h(x_i)$: kontribusi dari model tambahan h terhadap prediksi model ansambel saat ini
   c. Optimasi langkah pembelajaran (γ_m)
      Menemukan nilai γ_m yang meminimalkan fungsi kerugian pada setiap iterasi m dengan menambahkan model tambahan $h_m(x_i)$ ke dalam ansambel.
      $$\gamma_m = arg\ \min_{\gamma} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$
      Keterangan:
      γ_m : Langkah pembelajaran yang akan dievaluasi pada iterasi ke-m
      $h_m(x_i)$: Prediksi dari model tambahan h pada sampel (x_i) pada iterasi ke-m

d. Perbarui model

Memperbarui prediksi model ansambel dengan menggabungkan kontribusi dari model tambahan yang baru saja dipilih.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Keterangan:

$F_m(x)$: Prediksi dari model ansambel setelah iterasi ke-m.

$F_{m-1}(x)$: Prediksi model pada iterasi sebelumnya $m - i$.

Dengan demikian, langkah-langkah ini diulang secara berulang sampai kriteria penghentian tercapai atau sejumlah iterasi yang ditentukan tercapai, di mana setiap iterasi bertujuan untuk meningkatkan performa model ansambel dengan menambahkan model tambahan yang mampu mengoreksi kesalahan yang ada.

1.6.14 *Confussion Matrix*

Confussion Matrix adalah alat evaluasi yang digunakan dalam pengukuran performa suatu model klasifikasi. Matriks ini memberikan gambaran detail tentang seberapa baik suatu model dapat mengklasifikasikan *instance* data ke dalam kategori yang benar atau salah. *Confusion Matrix* biasanya digunakan dalam konteks pengujian atau validasi model klasifikasi, terutama ketika model tersebut memprediksi label atau kategori yang berbeda. *Confusion Matrix* memiliki empat variabel dalam proses klasifikasi, yaitu (Sanjaya et al., 2023):

Tabel 1. 1 *Confussion Matrix Plotting Classification*

		Predicted Class	
		Positive	***Negative***
Actual Class	***Positive***	*True Positive* (TP)	*False Positive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

Informasi yang diperoleh dari *True Positive* (TP) dan *True Negative* (TN) memberikan gambaran ketika pengklasifikasi berhasil mengenali dengan benar, sedangkan *False Positive* (FP) dan *False Negative* (FN) memberikan indikasi ketika terjadi kesalahan oleh pengklasifikasi (Ha et al., 2011). Dengan menggunakan nilai-nilai ini, berbagai metrik evaluasi dapat dihitung:

1. Akurasi (*Accuracy*): Persentase prediksi yang benar dari total prediksi yang dibuat oleh model untuk mengukur seberapa baik model dalam melakukan prediksi dengan benar secara keseluruhan.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1.4)$$

2. Presisi (*Precision*): Persentase prediksi positif yang benar dari semua prediksi yang dinyatakan positif oleh model untuk mengukur seberapa akurat model dalam mengidentifikasi positif yang sebenarnya dari semua prediksi positif yang dibuat.

$$\text{Presisi} = \frac{TP}{TP + FP} \times 100\% \quad (1.5)$$

3. *Recall* (*Sensitivity* atau *True Positive* Rate): Persentase prediksi positif yang benar dari semua kasus yang benar-benar positif untuk mengukur seberapa baik model dalam menangkap semua kasus positif yang sebenarnya.

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (1.6)$$

4. *F1-score*: Rata-rata harmonis dari presisi dan *recall* untuk memberikan satu angka yang memberikan gambaran seimbang tentang kinerja model antara presisi dan *recall*.

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Presisi} \times Recall}{\text{Presisi} + Recall} \times 100\% \quad (1.7)$$

Penggunaan *Confusion Matrix* sangat penting karena dapat memberikan wawasan yang lebih mendalam tentang kelemahan atau keunggulan suatu model klasifikasi. Dengan mengevaluasi berbagai metrik ini, pengembang atau peneliti dapat membuat penyesuaian atau perbaikan pada model mereka untuk mencapai performa yang lebih baik. *Confusion Matrix* juga membantu dalam menentukan apakah model cenderung melakukan kesalahan tertentu, seperti lebih sering mengklasifikasikan data sebagai salah satu kelas tertentu. Dengan demikian, *Confusion Matrix* memberikan landasan yang kuat untuk pengembangan dan peningkatan model klasifikasi.

BAB II

METODE PENELITIAN

2.1 Jenis Penelitian

Penelitian ini menggunakan jenis penelitian kuantitatif dengan pendekatan deskriptif untuk menganalisis sentimen pengguna X (twitter) terhadap kedatangan etnis Rohingya ke Indonesia. Jenis penelitian kuantitatif dipilih untuk memberikan gambaran yang sistematis dan objektif dalam mengumpulkan serta menganalisis data secara statistik. Sedangkan pendekatan deskriptif digunakan untuk merinci dan menganalisis data dengan cermat, memberikan pemahaman mendalam tentang pola sentimen yang muncul di kalangan pengguna X. Dengan demikian, penggunaan jenis penelitian kuantitatif dengan pendekatan deskriptif dalam penelitian ini diharapkan dapat memberikan pemahaman yang komprehensif mengenai sentimen pengguna X (twitter) terhadap isu kedatangan etnis Rohingya ke Indonesia, dengan hasil yang dapat diinterpretasikan secara luas dan mendukung pengembangan wawasan serta kebijakan yang relevan.

2.2 Sumber Data

Data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh dari media sosial X (twitter). Penggunaan data dari X sebagai sumber utama merupakan cara yang efektif untuk mendapatkan pandangan langsung dari masyarakat terkait isu kontemporer.

2.3 Objek Penelitian

Objek dalam penelitian ini adalah unggahan pengguna media sosial X (twitter) terhadap peristiwa kedatangan etnis Rohingya ke Indonesia. Dalam konteks ini, objek penelitian mencakup seluruh rangkaian interaksi dan respon yang diekspresikan oleh pengguna X (twitter) terkait isu tersebut.

2.4 Variabel Penelitian

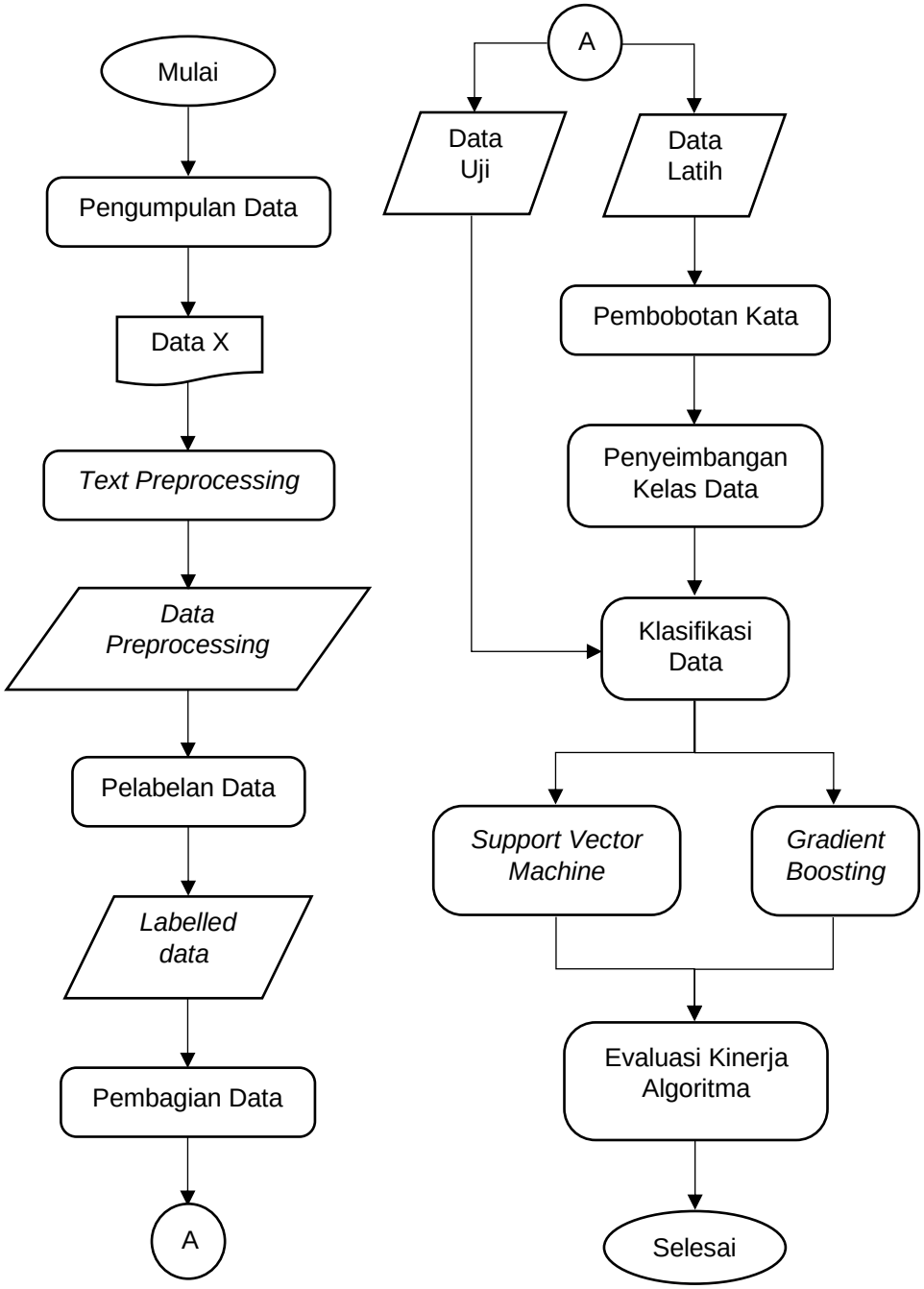
2.4.1 Variabel Independen

Variabel independen adalah variabel yang dianggap sebagai penyebab atau pemicu perubahan pada variabel lain dalam suatu penelitian. Dalam hal ini, teks yang dihasilkan oleh pengguna X (twitter) terkait isu kedatangan etnis Rohingya ke Indonesia.

2.4.2 Variabel Dependen

Variabel dependen adalah variabel yang dipengaruhi atau yang nilainya bergantung pada variabel independen. Dalam hal ini, hasil klasifikasi sentimen dari teks *tweet* menggunakan Algoritma *Support Vector Machine* dan *Gradient Boosting* dimana sentimen dapat berupa positif, negatif, atau netral.

2.5 Alur Penelitian



Gambar 2. 1 Alur penelitian