BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi informasi dan komunikasi pada era digital telah membawa perubahan besar dalam cara manusia berinteraksi dan mengekspresikan ide maupun emosinya. Media sosial menjadi salah satu platform utama yang memungkinkan jutaan pengguna di seluruh dunia, termasuk Indonesia, untuk saling berkomunikasi, berbagi informasi, dan mengekspresikan perasaan secara *real time*. Aplikasi X merupakan salah satu platform media sosial yang populer di Indonesia, di mana pengguna menghasilkan data teks yang sangat melimpah dalam bentuk opini, keluhan, serta ekspresi berbagai emosi. Data tekstual ini menyimpan nilai informasi yang penting untuk dianalisis, terutama dalam memahami kecenderungan emosional pengguna yang dapat dimanfaatkan untuk banyak kepentingan seperti analisis opini publik, pengembangan layanan pelanggan, dan pemantauan isu sosial.

Dalam beberapa tahun terakhir, penelitian mengenai *emotion detection* atau deteksi emosi lebih banyak berfokus pada sinyal fisiologis (misalnya EEG) maupun ekspresi wajah dan suara. Sementara itu, kajian mengenai deteksi emosi berbasis teks berbahasa Indonesia masih relatif terbatas. Padahal, media sosial seperti X merupakan salah satu sumber data yang sangat kaya dengan teks yang mencerminkan emosi pengguna dalam bentuk opini, keluhan, dan ekspresi perasaan sehari-hari.

Berbagai penelitian sebelumnya telah menggunakan beragam algoritma pembelajaran mesin dengan berbagai teknik representasi fitur. *Naïve Bayes* memiliki kelebihan sederhana dan cepat dalam klasifikasi teks, namun kelemahannya adalah tingkat akurasi yang menurun ketika data memiliki banyak fitur yang saling bergantung (LUBIS, 2024). *Support Vector Machine* (SVM) dikenal memiliki akurasi tinggi pada data berdimensi besar, tetapi memerlukan waktu komputasi yang cukup lama dan parameterisasi yang kompleks (Arifin, Rahman, Rintyarna, & Daryanto, 2023). Selanjutnya, pendekatan *deep learning*

seperti *Convolutional Neural Network* (CNN) memiliki kemampuan untuk menangkap representasi konteks teks dengan baik, namun kelemahannya adalah kebutuhan akan jumlah data yang sangat besar serta sumber daya komputasi yang tinggi sehingga tidak selalu efisien untuk seluruh kasus penelitian (Aulia, 2024). Di sisi lain, metode *K-Nearest Neighbors* (KNN) menawarkan mekanisme klasifikasi yang intuitif berbasis kedekatan data, namun performanya dapat menurun pada data berdimensi tinggi dan jumlah data yang besar karena biaya komputasi dan sensitivitas terhadap pemilihan parameter k (Cover & Hart, 1967; Han et al., 2012; Guo et al., 2003).

Dalam konteks representasi fitur teks, terdapat dua pendekatan utama yang banyak digunakan, yaitu pendekatan berbasis frekuensi statistik dan pendekatan berbasis *word embedding*. *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan metode klasik berbasis frekuensi yang telah terbukti efektif dalam berbagai tugas klasifikasi teks karena kesederhanaannya, interpretabilitasnya yang tinggi, serta kemampuannya memberikan bobot yang tepat pada kata-kata yang representatif (Salton & Buckley, 1988). TF-IDF bekerja dengan memberikan nilai tinggi pada kata yang sering muncul dalam dokumen tertentu namun jarang di dokumen lain, sehingga mampu menangkap kata kunci yang spesifik untuk setiap kategori emosi. Di sisi lain, *Word2Vec* sebagai teknik *word embedding* mampu menangkap hubungan semantik antar kata dalam ruang vektor, namun memerlukan korpus yang besar dan proses pelatihan yang lebih kompleks (Mikolov et al., 2013).

Berbeda dengan metode-metode tersebut, *Random Forest* memiliki kelebihan berupa kemampuannya menangani data berdimensi tinggi, relatif tahan terhadap *overfitting* karena menggunakan mekanisme *ensemble learning*, dapat menghasilkan klasifikasi yang stabil meskipun data mengandung *noise*, serta memiliki fleksibilitas tinggi dalam bekerja dengan berbagai jenis representasi fitur (Mahardika, 2024; Breiman, 2001). Karakteristik ini menjadikan *Random Forest* sangat cocok untuk dikombinasikan dengan metode representasi fitur berbasis frekuensi seperti TF-IDF, yang menghasilkan matriks fitur dengan dimensi tinggi namun *sparse*. Akan tetapi, *Random Forest* juga memiliki kelemahan, yaitu model yang terbentuk relatif sulit untuk diinterpretasikan karena kompleksitasnya, serta memerlukan tuning parameter yang tepat untuk mencapai performa optimal.

Berbagai penelitian sebelumnya menunjukkan bahwa algoritma klasifikasi tradisional seperti *Support Vector Machine* (SVM) dan *K-Nearest Neighbors* (KNN) mampu memberikan performa yang kompetitif pada tugas klasifikasi teks, termasuk analisis sentimen dan deteksi emosi, terutama ketika dipadukan dengan teknik representasi fitur yang tepat. Demikian pula, teknik *word embedding* seperti *Word2Vec* telah terbukti efektif dalam menangkap relasi semantik antar kata untuk berbagai aplikasi NLP (Mikolov et al., 2013). Hal ini menjadikan SVM, KNN, dan *Word2Vec* relevan untuk dijadikan model pembanding (*baseline*) dalam penelitian ini, karena ketiganya telah banyak digunakan sebagai standar *baseline* dalam studi-studi serupa dan terbukti mampu menghasilkan tingkat akurasi yang baik pada data teks berbahasa Indonesia.

Di sisi lain, karakteristik data teks dari media sosial X cenderung berdimensi tinggi, mengandung banyak *noise*, memiliki variasi kosakata yang luas, serta seringkali mengandung kata-kata informal dan tidak baku. Dalam konteks tersebut, kombinasi *Random Forest* dengan TF-IDF dipandang sangat cocok karena TF-IDF mampu memberikan representasi yang *robust* terhadap variasi kata melalui pembobotan statistik, sementara *Random Forest* memiliki kemampuan menangani matriks fitur berdimensi tinggi yang dihasilkan TF-IDF, relatif tahan terhadap *overfitting* melalui mekanisme *ensemble*, dan mampu menghasilkan klasifikasi yang stabil pada data dengan kualitas yang bervariasi. Dengan demikian, pemilihan *Random Forest* dan TF-IDF pada penelitian ini tidak hanya didasarkan pada *gap* penelitian, tetapi juga pada kesesuaian karakteristik kedua metode tersebut dengan sifat data teks media sosial yang digunakan.

Meskipun kombinasi *Random Forest* dan TF-IDF telah terbukti efektif dalam berbagai studi klasifikasi teks, penerapannya secara spesifik untuk kasus deteksi emosi pada teks berbahasa Indonesia dari media sosial X masih sangat terbatas. Fakta ini terlihat dari minimnya literatur akademik yang secara khusus mengeksplorasi efektivitas kombinasi *Random Forest* dengan TF-IDF pada dataset berbahasa Indonesia yang bersumber dari platform media sosial dengan karakteristik teks informal dan beragam. Oleh karena itu, agar pemilihan *Random Forest* dengan TF-IDF sebagai metode utama memiliki dasar empiris yang kuat, penelitian ini juga menyertakan kombinasi algoritma lain (SVM dan KNN dengan

TF-IDF) serta pendekatan *word embedding* (Word2Vec dengan ketiga algoritma) sebagai model *baseline* untuk perbandingan.

Pada sisi representasi fitur, penelitian ini memanfaatkan dua pendekatan yang berbeda, yaitu *Term Frequency–Inverse Document Frequency* (TF-IDF) sebagai metode pembobotan klasik berbasis frekuensi kata yang menjadi fokus utama penelitian, serta *Word2Vec* sebagai teknik *word embedding* kontemporer yang mampu menangkap hubungan semantik antar kata dalam ruang vektor yang digunakan sebagai pembanding. Kombinasi tiga algoritma (*Random Forest*, SVM, dan KNN) dengan dua jenis representasi fitur (TF-IDF dan *Word2Vec*) menghasilkan enam model klasifikasi emosi yang akan dievaluasi. Enam model ini digunakan untuk melakukan perbandingan kinerja secara sistematis, dengan *Random Forest* + TF-IDF sebagai model utama yang menjadi fokus analisis mendalam, sedangkan kombinasi lainnya berfungsi sebagai *baseline* untuk menunjukkan keunggulan komparatif dari model utama.

Penelitian ini akan memetakan teks berbahasa Indonesia yang diambil dari media sosial X ke dalam beberapa kategori emosi sesuai dengan label pada dataset yang digunakan, seperti misalnya emosi senang, sedih, marah, takut, dan kategori emosi lain yang relevan dengan konteks percakapan pengguna di platform tersebut. Hasil klasifikasi emosi tersebut diharapkan dapat memberikan gambaran lebih rinci mengenai distribusi dan kecenderungan emosi pengguna, sehingga tidak hanya berhenti pada polarisasi positif-negatif, tetapi mampu mengungkap variasi emosi yang lebih spesifik.

Pemilihan TF-IDF sebagai representasi fitur utama dalam penelitian ini didasarkan pada beberapa pertimbangan. Pertama, TF-IDF memiliki interpretabilitas yang tinggi karena bobot setiap kata dapat dijelaskan secara matematis melalui frekuensi kemunculan dan distribusinya dalam korpus. Kedua, TF-IDF telah terbukti efektif dan efisien untuk tugas klasifikasi teks pada berbagai bahasa, termasuk bahasa Indonesia, tanpa memerlukan proses pelatihan tambahan seperti halnya *word embedding*. Ketiga, TF-IDF menghasilkan representasi yang *sparse* namun informatif, yang sangat sesuai dengan karakteristik *Random Forest* dalam menangani data berdimensi tinggi. Keempat, kombinasi TF-IDF dan

Random Forest memiliki keunggulan komputasional dibandingkan pendekatan *deep learning* atau *word embedding* yang memerlukan sumber daya lebih besar. Kelima, TF-IDF mampu menangkap kata-kata kunci spesifik yang menjadi penanda emosi dalam teks berbahasa Indonesia, seperti kata sifat dan kata keterangan emosional yang memiliki distribusi unik pada setiap kategori emosi.

Seiring dengan meningkatnya penggunaan media sosial di Indonesia, dibutuhkan sebuah sistem yang mampu melakukan deteksi emosi pada teks berbahasa Indonesia secara lebih akurat. Sistem tersebut penting untuk mendukung berbagai kebutuhan, seperti analisis opini publik, peningkatan kualitas layanan, hingga pemetaan isu sosial. Output deteksi emosi yang dihasilkan dapat dimanfaatkan dalam berbagai kebutuhan praktis, antara lain sebagai dasar analisis opini publik yang lebih mendalam, bahan pertimbangan bagi pengambil kebijakan dalam memantau isu sosial yang sensitif, serta sebagai masukan bagi pengembang layanan untuk meningkatkan kualitas interaksi dan respons terhadap pengguna.

Secara teoretis, penelitian dalam bidang ini seharusnya mengintegrasikan algoritma pembelajaran mesin yang tepat guna untuk meningkatkan akurasi klasifikasi emosi. Dengan adanya model utama *Random Forest* dan dua model *baseline* SVM dan KNN, penelitian ini diharapkan mampu menjawab apakah Random Forest benar-benar memberikan peningkatan kinerja yang signifikan dibandingkan algoritma yang selama ini banyak digunakan, sekaligus mengidentifikasi kombinasi algoritma dan representasi fitur yang paling efektif untuk deteksi emosi pada teks berbahasa Indonesia di platform X.

Dengan demikian, penelitian ini tidak hanya mengisi kekosongan dalam literatur akademik yang berkaitan dengan penggunaan kombinasi *Random Forest* dan TF-IDF untuk deteksi emosi teks berbahasa Indonesia dari media sosial, tetapi juga memberikan kontribusi empiris mengenai perbandingan pendekatan representasi fitur (TF-IDF vs *Word2Vec*) serta algoritma klasifikasi (*Random Forest* vs SVM dan KNN) dalam konteks bahasa Indonesia. Selain itu, penelitian ini mendorong pengembangan aplikasi kecerdasan buatan yang lebih efisien, *interpretabel*, dan adaptif terhadap kebutuhan serta karakteristik bahasa dan budaya Indonesia di era digital saat ini.

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang, maka dapat diidentifikasi beberapa permasalahan penelitian sebagai berikut:

1. Metode yang banyak digunakan dalam penelitian sebelumnya masih didominasi oleh algoritma *Naïve Bayes*, *Support Vector Machine* (SVM), serta pendekatan *deep learning* dengan *word embedding*, sedangkan penggunaan kombinasi *Random Forest* dengan TF-IDF untuk kasus deteksi emosi teks berbahasa Indonesia dari media sosial belum banyak dieksplorasi secara mendalam.
2. Potensi pemanfaatan data teks dari media sosial, khususnya X, yang memiliki jumlah data sangat besar, belum dimanfaatkan secara maksimal untuk pengembangan sistem deteksi emosi.
3. Belum terdapat kajian yang mendalam mengenai kinerja kombinasi *Random Forest* dengan TF-IDF dalam mengklasifikasikan emosi pada teks berbahasa Indonesia dari media sosial, khususnya dalam perbandingan dengan pendekatan *word embedding* dan algoritma klasifikasi lainnya, dengan pengukuran parameter evaluasi seperti akurasi, presisi, *recall*, dan *f1-score*.
4. Sistem deteksi emosi berbasis teks yang ada saat ini belum sepenuhnya mendukung kebutuhan praktis, misalnya dalam analisis opini publik, peningkatan kualitas layanan, maupun pemetaan isu sosial yang relevan di konteks Indonesia.

1.3 Pembatasan Masalah

Agar penelitian ini lebih terarah dan sesuai dengan tujuan yang hendak dicapai, maka peneliti memberikan pembatasan masalah sebagai berikut:

1. Penelitian ini hanya berfokus pada deteksi emosi berbasis teks dengan menggunakan data berbahasa Indonesia yang diambil dari media sosial X melalui *Apify*.
2. Jumlah data yang digunakan dalam penelitian ini dibatasi sebanyak 7.687 *tweet* berbahasa Indonesia.
3. *Tweet* yang diambil dibatasi pada rentang waktu tiga tahun terakhir dari saat pengambilan data.

4. Teks yang digunakan terbatas pada hasil *Pre-processing* data berupa *cleaning*, *case folding*, *stopword removal*, *tokenizing*, dan *stemming*, kemudian direpresentasikan menggunakan TF-IDF sebagai metode utama dan *Word2Vec* sebagai metode pembanding untuk mengubah data teks menjadi representasi numerik.
5. Algoritma yang digunakan dalam proses klasifikasi emosi dibatasi pada *Random Forest* sebagai metode utama yang dikombinasikan dengan TF-IDF, serta *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Word2Vec* (dikombinasikan dengan ketiga algoritma) sebagai *baseline* untuk perbandingan.
6. Penelitian ini hanya menganalisis performa sistem berdasarkan parameter evaluasi umum pada klasifikasi teks, yaitu akurasi, presisi, *recall*, dan *f1-score*.
7. Emosi yang dianalisis dibatasi pada kategori tertentu sesuai dengan label dataset yang diperoleh, dan tidak mencakup seluruh spektrum emosi manusia secara umum.

1.4 Perumusan Masalah

Berdasarkan latar belakang, identifikasi masalah, dan pembatasan masalah yang telah diuraikan, maka perumusan masalah dalam penelitian ini adalah bagaimana kinerja kombinasi algoritma *Random Forest* dengan representasi fitur TF-IDF dalam mendeteksi emosi pada teks berbahasa Indonesia yang diperoleh dari media sosial X dibandingkan dengan metode *baseline* lainnya?

1.5 Tujuan Penelitian

Tujuan penelitian ini adalah untuk menganalisis dan mengukur kinerja serta efektivitas kombinasi algoritma *Random Forest* dengan representasi fitur TF-IDF dalam klasifikasi emosi berbasis teks berbahasa Indonesia, serta membandingkannya dengan metode *baseline* yang menggunakan algoritma dan representasi fitur lainnya.

1.6 Manfaat Penelitian

Manfaat penelitian ini dibagi menjadi dua, yaitu manfaat teoritis yang berkaitan dengan kontribusi penelitian terhadap pengembangan ilmu pengetahuan, serta

manfaat praktis yang memberikan dampak langsung bagi pihak-pihak terkait dalam penerapan hasil penelitian.

1.6.1 Manfaat Teoritis

1. Memberikan kontribusi pada pengembangan ilmu *Natural Language Processing* (NLP) khususnya dalam deteksi emosi berbasis teks berbahasa Indonesia melalui eksplorasi kombinasi *Random Forest* dengan TF-IDF dan perbandingannya dengan pendekatan *word embedding*.
2. Memperkaya literatur terkait penerapan algoritma *Random Forest* dengan representasi fitur TF-IDF, serta memberikan wawasan komparatif terhadap teknik *Word2Vec* dalam konteks analisis teks media sosial berbahasa Indonesia.

1.6.2 Manfaat Praktis

Bagi Bidang Ilmu Komputer dan Prodi Pendidikan Teknik Informatika dan Komputer UNJ:

1. Menyediakan alat evaluasi objektif untuk menganalisis ekspresi emosi dalam data teks berbahasa Indonesia dari platform media sosial X.
2. Memfasilitasi pemetaan dan pengembangan model klasifikasi emosi yang sesuai dengan karakteristik bahasa Indonesia.

Bagi Dosen dan Peneliti:

1. Memberikan *insight* untuk pengembangan model *machine learning* berbasis data bahasa Indonesia.
2. Mendukung pengembangan penelitian lanjutan di bidang analisis sentimen dan deteksi emosi berbahasa Indonesia.
3. Mempermudah identifikasi peluang kolaborasi antar bidang penelitian terkait NLP dan kecerdasan buatan.

Bagi Mahasiswa:

1. Meningkatkan pemahaman praktis tentang implementasi algoritma *Random Forest* dan TF-IDF dalam klasifikasi teks.

2. Menjadi referensi aplikasi *machine learning* pada data teks media sosial yang relevan dengan konteks bahasa dan budaya Indonesia.
3. Membantu dalam pengembangan skripsi atau penelitian lanjutan pada bidang *machine learning* dan NLP.