

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, berita dan informasi berkembang serta menyebar dengan sangat cepat. Seiring berjalannya waktu, kebutuhan masyarakat akan informasi semakin banyak. Dari perkembangan teknologi informasi dan komunikasi masyarakat dapat memperoleh berbagai informasi dengan lebih cepat dan mudah. Layanan informasi sudah banyak tersedia untuk masyarakat pengguna sosial media seperti Instagram yang saat ini sering dipakai untuk menyampaikan berita terkait isu politik, ekonomi, sosial, budaya, kriminal, gaya hidup, dan lain-lain. Serta menjadi wadah untuk menyalurkan opini sekaligus menjadi ruang diskusi untuk publik.

Akun informasi sangat banyak tersedia di platform Instagram, namun ada beberapa akun yang menyajikan informasi seperti akun Instagram Folkative. Berdasarkan data *SocialBlade* (2025), akun Folkative memiliki lebih dari 6,7 juta pengikut pada kategori akun sosial, budaya, dan tren *digital*, menjadi salah satu akun informasi paling populer di negara Indonesia. Komentar-komentar yang ditulis dalam beberapa unggahan mengandung respon positif, negatif, dan netral terhadap isu yang dibahas. Namun komentar atau opini publik yang dituangkan sangat beragam dan sering bersifat ambigu, banyak kata singkat, kesalahan dalam mengetik dan dipengaruhi oleh konteks budaya lokal (bahasa gaul).

Dalam hal ini, analisis dibutuhkan dalam membantu memahami opini publik pada komentar akun Folkative. “Analisis sentimen merupakan metode evaluasi opini dan respon dalam teks” (Sjoraida et al., 2024). Melalui teknik ini, peneliti dapat mengidentifikasi sebuah komentar bersifat positif dan negatif terhadap sebuah isu. Menganalisis sentimen dapat menghasilkan data yang lebih objektif tentang pendapat masyarakat mengenai isu sosial yang sedang ramai dibicarakan. Memahami pola sentimen masyarakat penting untuk mengetahui secara jelas mengenai pemahaman masyarakat terhadap isu-isu sosial sedang populer di sosial media.

Namun, tantangan dalam sebuah analisis sentimen berada pada kerumitan bahasa yang natural. Terutama pada kalimat pendek seperti komentar pengguna Instagram yang informal, bahasa campuran, kesalahan dalam mengetik, serta banyak mengandung sarkasme. Model algoritma klasik seperti *Support Vector Machine* (SVM) sering digunakan karena performanya yang bagus dalam tugas klasifikasi, akan tetapi performa klasifikasi sangat dipengaruhi oleh kualitas representasi fitur yang digunakan dalam proses ekstraksi teks. Berdasarkan hasil dari penelitian sebelumnya yang dilakukan oleh Luqyana, Cholissodin, dan Perdana (Luqyana et al., 2018), berjudul “Analisis Sentimen *Cyberbullying* pada Komentar Instagram dengan Metode Klasifikasi *Support Vector Machine* (SVM)” memberikan hasil akurasi terbaik yaitu 90% dalam menganalisis sentimen komentar di Instagram. Saat menerapkan model tradisional seperti *Bag of Words* (BoW) atau TF-IDF belum mampu memahami hubungan semantik antar kata karena hanya memperhatikan frekuensi kata, sehingga hasil klasifikasi kurang maksimal (Dang et al., 2020). Meskipun demikian, TF-IDF masih sering diterapkan dalam analisis sentimen karena memiliki proses representasi yang mudah dan efisien pada berbagai macam data teks, terutama pada komentar media sosial yang cenderung mengandung kata-kata sentimen secara eksplisit. Oleh karena itu, TF-IDF tetap digunakan sebagai metode *baseline* untuk dibandingkan dengan representasi berbasis *word embedding*.

Untuk merepresentasikan fitur teks, kini dikembangkan pendekatan *word embedding*. Metode ini berfungsi mengubah kata menjadi bentuk vector numerik agar hubungan semantik antar kata dapat dipahami oleh system. Dua metode yang paling populer *FastText* dan BERT (*Bidirectional Encoder Representations from Transformers*), terbukti efisien dan sudah banyak digunakan penelitian pemrosesan bahasa alami (Bag of Tricks for Efficient Text Classification - ACL Anthology, n.d.; Joulin et al., 2017). “*FastText* dikembangkan oleh Facebook AI Research” (Bojanowski et al., 2017), unggul dalam penanganan kata *out-of-vocabulary* karena memanfaatkan informasi *subword* (*character n-gram*). Pendekatan ini memungkinkan model untuk membangun representasi dari kata baru dengan menggabungkan vektor dari sub-kata penyusunnya, sehingga membuatnya sangat kuat terhadap kesalahan seperti kata-kata informal, salah ketik, variasi yang

kompleks yang lazim ditemui di media sosial (Joulin et al., 2017b). “Dalam penggunaan BERT (*Bidirectional Encoder Representations from Transformers*), teknik ini memanfaatkan representasi bahasa yang diperoleh dari *pretraining deep bidirectional representation* untuk memahami konteks kiri dan kanan dalam teks” (Wardhana & Purwitasari, 2019). Arsitektur ini memproses seluruh urutan kata secara bersamaan dari kedua arah (kiri dan kanan), sehingga makna vektor sebuah kata dapat berubah tergantung pada seluruh kata lain dalam kalimatnya. Kemampuan ini memungkinkan BERT membedakan perbedaan makna yang halus dan kata-kata dengan banyak arti, hal yang sebelumnya tidak dapat dipecahkan oleh model *embedding* yang biasa (Qiu et al., 2020).

Berdasarkan uraian tersebut, penelitian ini berjudul “Analisis Perbandingan Representasi Fitur TF-IDF, *FastText*, dan BERT Pada Model *Support Vector Machine* Untuk Klasifikasi Sentimen Komentar Instagram Folkative”, bertujuan menerapkan dan membandingkan 3 metode representasi fitur, yaitu TF-IDF sebagai *baseline*, dan dua metode *word embedding* *FastText*, dan BERT (*Bidirectional Encoder Representations from Transformers*) pada model *Support Vector Machine* (SVM) untuk klasifikasi sentimen komentar Instagram Folkative. Penelitian ini juga bertujuan menganalisis performa pada masing-masing representasi fitur berdasarkan nilai *accuracy*, *precision*, *recall*, *f1-score*, dan pengaruh representasi fitur terhadap hasil klasifikasi sentimen.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, penulis merumuskan topik penelitian ini, yaitu:

1. Bagaimana penerapan metode representasi fitur TF-IDF, *FastText*, dan BERT pada model *Support Vector Machine* (SVM)?
2. Bagaimana hasil perbandingan performa metode representasi fitur TF-IDF, *FastText*, dan BERT pada model *Support Vector Machine* (SVM) berdasarkan nilai *accuracy*, *precision*, *recall*, dan *f1-score*?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dirumuskan, akan diperoleh hasil yang sesuai dengan tujuan penelitian, yaitu:

1. Menerapkan metode representasi fitur TF-IDF, *FastText*, dan BERT pada model *Support Vector Machine* (SVM).
2. Menganalisis dan membandingkan tiga representasi fitur TF-IDF, *FastText*, dan BERT pada model *Support Vector Machine* (SVM) berdasarkan nilai *accuracy*, *precision*, *recall*, dan *f1-score*.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik dari segi teori maupun praktik, yang akan diuraikan sebagai berikut:

1. Penelitian ini menunjukkan pengaruh representasi fitur terhadap performa klasifikasi sentimen menggunakan model *Support Vector Machine* (SVM).
2. Hasil penelitian dapat digunakan sebagai acuan untuk peneliti selanjutnya dalam analisis sentimen komentar media sosial yang bersifat tidak terstruktur.
3. Memberikan wawasan bagi praktisi data atau pengelola akun media sosial untuk memahami pandangan dan respon publik.

1.5 Batasan Masalah

Penulis menerapkan beberapa pembatasan masalah dalam penelitian ini untuk memastikan analisis yang fokus dan obyektif, antara lain:

1. Penelitian ini menggunakan data komentar pengguna dari akun Instagram Folkative yang diambil dari beberapa unggahan.
2. Penelitian ini dibatasi pada komentar akun Instagram Folkative yang dan digunakan sebagai dataset dalam proses analisis sentimen.
3. Penelitian ini menggunakan algoritma *Support Vector Machine* (SVM) sebagai model utama untuk melakukan analisis sentimen
4. Penelitian ini menggunakan TF-IDF sebagai *baseline* dan *word embedding* (*FastText* & BERT) sebagai representasi fitur, yang kemudian dibandingkan performanya.
5. Klasifikasi sentimen hanya dibatasi untuk 2 kategori (positif dan negatif)

6. Proses pelabelan dilakukan secara semi-automatis yaitu pelabelan awal dilakukan secara otomatis, kemudian divalidasi oleh pakar ahli bahasa Indonesia.
7. Penelitian ini berfokus pada analisis komentar tanpa mempertimbangkan aspek lain seperti jumlah *like*, waktu komentar yang diunggah, *caption* postingan, dan lain-lain.
8. Implementasi model klasifikasi sentimen menggunakan bahasa pemrograman *Python* pada *platform Google Colab*.

