BAB I
PENDAHULUAN

1.1 Latar Belakang

Sarkasme menghasilkan perbedaan arti harfiah dari kata-kata yang digunakan dan maksud yang sebenarnya ingin disampaikan [1], [2]. Sarkasme adalah bentuk sindiran yang kompleks, karena ekspresi sindiran sering kali bersifat tersirat biasanya digunakan untuk menyampaikan kritik dengan cara yang menyakitkan atau mengejek [3], [4]. Sarkasme sering ditemukan diantara komentar-komentar orang pada media sosial untuk mengekspresikan perasaan dan opini mereka. Komentar yang terdapat pada media sosial dapat digunakan sebagai data sentimen yang berpotensi dimanfaatkan untuk berbagai keperluan. Salah satu manfaat utama dari data ini adalah memahami sentimen masyarakat (positif, negatif atau netral) terhadap berbagai topik, seperti produk, film, politik, dan isu-isu terkini guna membantu dalam memodelkan tren dan strategi bisnis [4] - [7]. Kemampuan untuk mengenali sarkasme sangat penting karena dapat memengaruhi hasil analisis karena sarkasme dapat menyebabkan kebisingan (*noise*) pada data [7]. Upaya untuk membuat model deteksi sarkasme guna meningkatkan keakuratan analisis sentimen pada data media sosial telah dilakukan oleh beberapa peneliti.

Debby & Auliya (2020) mengusulkan *support vector machine* (SVM) untuk analisis sentimen dan *random forest classifier* untuk deteksi sarkasme [4]. Hasil penelitian menunjukkan nilai akurasi sebesar 77,79%, presisi 64,01%, *recall* 62,45%, dan F1-*score* 62,32%. Penelitian Liu et al (2024) berkontribusi pada kinerja prediktif keseluruhan dari model SAHFN-RoBERTa dengan akurasi 92,73% dan skor F1 92,63%. Namun terdapat keterbatasan dataset yang memungkinkan model tidak dapat secara efektif menggeneralisasi hasilnya ke situasi yang berbeda [3]. Lalu pada *literatur review* yang dilakukan oleh Jihad & Ilavarasan (2020), terlihat bahwa deteksi sarkasme menggunakan *bidirectional encoder representations from transformers* (BERT) mencapai F1-*score* tertinggi sebesar 92,4% [1]. MD Saifullah et al (2023) pada penelitiannya melakukan deteksi sarkasme menggunakan *deep learning* dan hasil evaluasi menunjukkan regresi logistik menghasilkan 94% untuk akurasi, F1-*score*, *recall*, dan 95% untuk nilai presisi [8]. Ramisa Anan et al [9] melakukan penelitian untuk mendeteksi sarkasme dalam teks bahasa Bangla dengan mengusulkan model berbasis BERT dan menerapkan teknik gabungan untuk menjelaskan keputusan model menggunakan *Local Interpretable Modelagnostic Explanations* (LIME). Model yang dihasilkan mencapai tingkat akurasi yang sangat tinggi, yaitu 99,60% [9]. Berdasarkan

penelitian-penelitian yang membahas penggunaan BERT, penulis melihat bahwa BERT dapat berpotensi meningkatkan akurasi pada model deteksi sarkasme yang telah dilakukan oleh Debby & Auliya (2020).

BERT adalah model pemrosesan bahasa alami yang menggunakan arsitektur transformer [10], [11]. BERT menggunakan representasi yang bidireksional dan mempertimbangkan konteks kata dalam teks dari dua arah. Hal ini memungkinkan BERT lebih baik untuk memahami arti dan konteks kata-kata. Dengan representasi bidireksional ini, BERT dapat mengatasi berbagai tantangan dalam NLP, seperti pemahaman konteks kompleks, penafsiran kalimat yang ambigu, dan pengenalan pola yang rumit dalam teks, sehingga menjadikannya model yang sangat efektif untuk berbagai tugas pemrosesan bahasa alami [12], [13]. Adapun jenis-jenis BERT antara lain XLM-R. ArBERT, AraBERT, IndoBERT, AlBERT, RoBERTa, IndoBERT, dan lain sebagainya [14]. Penelitian ini akan menggunakan IndoBERT untuk deteksi sarkasme. IndoBERT adalah model bahasa alami yang dirancang khusus untuk bahasa Indonesia [10]. Pada penelitian yang dilakukan oleh Sani et al (2022), IndoBERT digunakan untuk melakukan deteksi *Indonesian fake news* dan menghasilkan akurasi sebesar 94,66% [11].

Berdasarkan uraian di atas, maka penelitian ini dilakukan untuk mengidentifikasi serta mengungkap kalimat-kalimat yang bersifat sarkastik dalam teks berbahasa Indonesia, serta menguji pengaruh model IndoBERT terhadap kinerja *Random Forest Classifier* dalam mendeteksi sarkasme terhadap analisis sentimen. Sehingga penelitian ini diberi judul **"PENERAPAN MODEL INDOBERT DALAM DETEKSI SARKASME MENGGUNAKAN RANDOM FOREST PADA ANALISIS SENTIMEN"**.

1.2 Rumusan Masalah

Berdasarkan penjelasan latar belakang di atas, masalah yang akan diselesaikan pada penelitian ini adalah apakah model IndoBERT berpengaruh dalam meningkatkan akurasi model *Random Forest Classifier* pada deteksi sarkasme dalam teks berbahasa Indonesia.

1.3 Tujuan

Penelitian ini bertujuan untuk mengidentifikasi dan mengungkap sarkasme dalam teks berbahasa Indonesia dengan memanfaatkan hasil analisis sentimen. Selain itu, penelitian ini juga berkontribusi pada pengembangan model dan teknik dalam *Natural Language Processing* (NLP), terutama dalam penggunaan model berbasis bahasa Indonesia (IndoBERT).

1.4 Manfaat

Adapun manfaat yang diharapkan dari penelitian ini adalah:

1. Memberikan wawasan tentang bagaimana cara mendeteksi dan mengidentifikasi sarkasme sehingga dapat bermanfaat dalam berbagai bidang, termasuk analisis opini publik, deteksi emosi dalam percakapan *online*, dan pemahaman interaksi sosial.
2. Memberikan landasan bagi pengembangan lebih lanjut pada analisis sentimen dalam bahasa Indonesia.

1.5 Ruang Lingkup

Ruang lingkup yang ditentukan untuk menghindari perluasan dalam penelitian ini adalah sebagai berikut:

1. Pendekatan analisis sentimen menggunakan metode SVM.
2. Data yang digunakan untuk analisis sentimen berasal dari komentar pada 8 video YouTube yang membahas terkait "Dinasti Politik Jokowi" dan 7 video yang membahas "Menkominfo Indonesia" yang diunggah pada tahun 2023 dan tahun 2024.
3. Pemberian label kalimat sarkas ("*sarcastic*") dan kalimat bukan sarkas ("*not sarcastic*") menggunakan 3 ekstraksi fitur yaitu *semantic related*, *punctuation related*, *lexical and syntactic*.
4. Model IndoBERT akan digunakan sebagai *feature extractor* dalam proses deteksi sarkasme.
5. Pengujian akan dilakukan menggunakan 10-Fold *Cross Validation* dengan metrik akurasi, presisi, *recall*, dan F1-*score*.
6. Bahasa yang menjadi fokus analisis adalah Bahasa Indonesia.