1. Pendahuluan

Latar Belakang

COVID-19 (*Corona Virus Disease*) adalah penyakit menular yang disebabkan oleh virus Corona yang paling baru ditemukan. Virus baru ini tidak diketahui sebelum wabah dimulai di Wuhan, Cina, pada bulan Desember 2019 [1]. Penyebaran COVID-19 yang terjadi di Indonesia sangatlah cepat, hingga tanggal 17 Juni 2020 tercatat jumlah kasus terkonfirmasi sebanyak 41.431 dan kasus meninggal sebanyak 2.276 [2]. Hal tersebut membuat masyarakat menjadi resah dan menimbulkan berbagai komentar mengenai COVID-19. Banyak masyarakat Indonesia yang mengutarakan komentar mereka mengenai COVID-19 melalui sosial media Twitter. Komentar tersebut dapat dimanfaatkan untuk melakukan analisis sentimen guna mengetahui kecenderungan komentar masyarakat terhadap peristiwa pandemi COVID-19, apakah cenderung berkomentar positif, netral, ataupun negatif. Analisis sentimen merupakan penambangan opini yang digunakan untuk mengenali konten di *website*. Analisis sentimen bertujuan untuk menghasilkan ungkapan yang sebenarnya dari seseorang mengenai suatu produk, layanan, fim, berita, masalah tertentu dan sebagainya [3].

Dalam melakukan analisis sentimen menggunakan data Twitter, terdapat permasalahan yang akan dihadapi. Umumnya data *tweet* masih mengandung banyak kata yang tidak baku seperti penulisan kata yang disingkat dan pengggunaan bahasa gaul. Hal tersebut disebabkan karena Twitter memiliki batasan dalam penulisan dengan maksimal 140 karakter untuk sekali unggah *tweet*. Oleh karena itu, perlu dilakukan *preprocessing* terhadap data *tweet* sebagai tahapan awal dalam analisis sentimen sehingga menghasilkan bentuk data yang lebih baik yang dapat digunakan pada proses lainnya. *Preprocessing* berfungsi untuk menanggulangi kesalahan dalam mengambil ciri atau atribut dan dapat menurunkan performa analisis sentimen secara signifikan [4].

Penelitian mengenai *text preprocessing* pernah dilakukan oleh beberapa peneliti. Pada penelitian [5] meneliti pengaruh *preprocessing* terhadap performa analisis sentimen. Hasil penelitian menunjukkan bahwa akurasi klasifikasi mengalami peningkatan sebesar 20.4% dengan menggunakan *stopword removal*, *stemming*, dan *feature selection*. Kemudian penelitian [6] melakukan pengujian teknik *preprocessing* dalam klasifikasi. Hasil pengujian menunjukkan bahwa penggunaan teknik *preprocessing cleaning*, *case folding*, dan *stemming* tanpa menggunakan *stopword removal* mampu meningkatkan akurasi sistem, dengan perolehan akurasi sebesar 94.24%. Pada penelitian [7] melakukan pengujian teknik *preprocessing tokenization*, *stopword removal*, dan *stemming* pada tiga dataset yang berbeda. Hasil penelitian menunjukkan setelah menggunakan pemilihan fitur dan representasi yang tepat akurasi analisis sentimen dapat ditingkatkan. Menurut penelitian ini, *preprocessing* menjadi langkah yang penting dalam analisis sentimen, karena pemilihan teknik *preprocessing* yang tepat dapat meningkatkan kinerja klasifikasi.

Berdasarkan hasil penelitian [5, 6, 7], menunjukkan bahwa *preprocessing* memiliki pengaruh yang cukup baik dalam meningkatkan kinerja sistem. Namun, dari penelitian sebelumnya tidak membahas mengenai pengaruh dari berbagai teknik *preprocessing* yang digunakan dan juga belum diketahui kombinasi *preprocessing* seperti apa yang menghasilkan kinerja analisis sentimen yang optimal. Maka dari itu, penelitian ini akan berfokus pada penerapan berbagai teknik *preprocessing*, sehingga dapat diketahui pengaruh *preprocessing* terhadap kinerja analisis sentimen. Dan dapat diketahui kombinasi teknik *peprocessing* mana yang akan menghasilkan kinerja analisis sentimen komentar Twitter terbaik. Teknik *preprocessing* yang digunakan pada penelitian ini berdasarkan dari penelitian sebelumnya yaitu *case folding*, *cleaning*, *stopword removal*, *stemming* dan menambahkan teknik normalisasi kata untuk mengatasi adanya penggunaan kata-kata yang tidak baku yang sering muncul ketika menggunakan data Twitter.

Penelitian ini menggunakan metode klasifikasi Support Vector Machine (SVM), SVM dipilih karena menurut [8] SVM memberikan performa algoritma yang jauh lebih baik dibandingkan dengan Naïve Bayes. Proses pembobotan kata menggunakan TF-IDF. Ekstraksi fitur menggunakan N-Gram karena berdasarkan [9], menerapkan N-Gram dapat meningkatkan akurasi klasifikasi data *tweet*. Selain itu, penelitian ini menerapkan Mutual Information (MI) sebagai seleksi fitur karena dengan menerapkan seleksi fitur dapat meminimalkan *overfitting* yang dapat menghilangkan data redundan dan *noise* [10]. MI dipilih karena memiliki titik fokus terdahap hubungan *term* kata dengan suatu kelas, sehingga fitur yang dihasilkan mampu meningkatkan akurasi klasifikasi [6].

Topik dan Batasannya

Pada penelitian ini penulis akan membangun model untuk analisis sentimen *tweet* komentar mengenai COVID-19. Penelitian ini akan berfokus pada proses *preprocessing*, seleksi fitur, dan ekstraksi fitur. Pada proses *preprocessing*, penulis akan membandingkan penggunaan kombinasi teknik normalisasi kata, *cleaning*, *stopword removal*, dan *stemming*. Proses seleksi fitur menggunakan metode Mutual Information (MI). Pada proses ekstraksi fitur menggunakan N-Gram. Kemudian pada proses klasifikasi menggunakan Support Vector Machine (SVM). Pada penelitian ini terdapat beberapa batasan masalah, yaitu dataset COVID-19 yang digunakan bersumber dari Twitter dengan jumlah 1080 *tweet* yang diambil dalam rentang bulan Maret-April 2020 dan hanya

berfokus pada tweet berbahasa Indonesia. Dataset dilabelkan secara manual kedalam tiga kelas, yaitu positif, negatif, dan netral. Proses ekstraksi fitur menggunakan unigram dan bigram.

Tujuan

Penelitian ini dilakukan dengan tujuan untuk menganalisis pengaruh dari pengkombinasian teknik *preprocessing* dan untuk mengetahui kombinasi *preprocessing* yang dapat menghasilkan akurasi sistem yang optimal. Menganalisis pengaruh penggunaan seleksi fitur Mutual Information terhadap akurasi sistem. Serta menganalisis pengaruh penggunaan fitur unigram dan bigram terhadap akurasi dalam analisis sentimen *tweet* komentar mengenai COVID-19.

Organisasi Tulisan

Bagian selanjutnya pada penelitian ini adalah bagian 2 yang membahas mengenai studi terkait dengan penelitian yang dilakukan, bagian 3 membahas rancangan sistem yang dibangun, bagian 4 membahas evaluasi dari hasil pengujian, dan bagian 5 membahas kesimpulan dari penelitian ini dan saran untuk penelitian selanjutnya.