Pengaruh *Stemming* Bahasa Indonesia Terhadap Analisis Sentimen pada *Twitter* (Menggunakan Dataset: Gojek)

Indah Ayu Nur Fauziah[1], **Yuliant Sibaroni**[2], **Kemas Muslim Lhaksmana**[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]indahanf@students.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id,
[3]kemasmuslim@telkomuniversity.ac.id

1. Pendahuluan

Latar Belakang

Transportasi *online* merupakan peralihan pengguna kendaraan pribadi menjadi angkutan umum dalam kota dengan memanfaatkan teknologi *online* baik dalam pemesanan maupun transaksi. Sejak tahun 2015, masyarakat Indonesia sudah mulai menyukai pemanfaatan alat transportasi ini sebagai salah satu pilihan untuk berpergian atau memesan makanan. Pada penelitian ini, transportasi online akan digunakan sebagai dataset untuk analisis sentimen. Analisis sentimen merupakan salah satu cabang ilmu dari data mining untuk menganalisis, memahami, mengolah, dan mengekstrak data tekstual yang berupa opini terhadap entitas topik tertentu [4]. Bagian yang sangat penting pada analisis sentimen adalah *preprocessing*. *Preprocessing* merupakan proses untuk menanggulangi salah mengambil ciri atau atribut. Karena kesalahan dalam menggunakan ciri atau atribut dapat menurunkan performa analisis sentimen secara signifikan. Salah satu proses pada *preprocessing* adalah proses *stemming*, di mana *stemming* digunakan untuk meningkatkan performa *Information Retrieval* dengan mentransformasikan kata-kata dalam sebuah dokumen teks ke kata dasarnya.

Masalah yang seringkali ditemui dalam pengerjaan *stemming* adalah ambiguitas, *overstemming* dan *understemming*. *Word Sense Disambiguation* (WSD) adalah suatu proses mengidentifikasi makna kata yang memiliki sejumlah makna yang berbeda dalam suatu kalimat tertentu [7]. *Overstemming* adalah kata yang terlalu banyak dipotong setelah dilakukan proses *stemming* dibandingkan dengan jumlah kata dalam dokumen [5]. *Understemming* adalah kata yang terlalu sedikit dipotong setelah dilakukan proses *stemming* dibandingkan dengan jumlah kata dalam dokumen [5]. Menurut Tahitoe dan Purwitasari, algoritma ECS Stemming masih belum bisa menyelesaikan permasalahan *overstemming* dan *understemming* [6], begitupun algoritma *stemming* lainnya seperti yang telah diteliti oleh Simarangkir mengenai perbandingan algoritma *stemming* Nazief Adriani, algoritma Vega, algoritma Arifin Setiono, dan algoritma Tala dengan hasil algoritma terbaik dengan penggunaan kamus ialah algoritma Nazief Adriani [5].

Studi ini akan membandingkan antara tiga kondisi, yaitu proses analisis sentimen tanpa penggunaan *stemming* pada *processing*, analisis sentimen dengan *stemming* menggunakan algoritma Nazief Adriani dan Stemming Sastrawi. Pengukuran *stemming* untuk teks Bahasa Indonesia pada kedua algoritma tersebut belum pernah dilakukan sebelumnya. Studi ini membandingkan Stemming Sastrawi yang merupakan algoritma *stemming* terbaru [2] dengan algoritma *stemming* Nazief Adriani yang memiliki presisi lebih tinggi dibanding algoritma *stemming* lainnya [3]. Algoritma *stemming* Nazief Adriani juga dapat meningkatkan *recall* [1].

Topik yang dibahas pada tugas akhir ini adalah menganalisis pengaruh *stemming* Bahasa Indonesia terhadap analisis sentimen. Penelitian ini dilakukan karena belum ada yang khusus mempelajari pengaruh *stemming* Bahasa Indonesia untuk analisis sentimen pada *twitter* mengenai jasa transportasi *online*. Metode klasifikasi yang digunakan adalah *Support Vector Machine* (SVM). Pada penelitian ini terdapat beberapa batasan yakni dataset yang dianalisis berupa *tweet* berBahasa Indonesia dengan jumlah data tweet sebanyak 2500 *tweet* dan penelitian yang akan dilakukan penulis berfokus pada proses *stemming* dengan menggunakan algoritma Nazief Adriani, dan Sastrawi, untuk proses ekstraksi fitur akan menggukan *N-gram*, TF-IDF untuk proses pembobotan, dan proses klasifikasi untuk bagian *learning* akan menggunakan algoritma SVM.

Penelitian ini berfokus pada analisis sentimen terhadap pengaruh *stemming* Bahasa Indonesia pada *twitter* dengan dataset gojek dan untuk mengetahui penanganan *word sense disambiguation*, *overstemming* dan *understemming* pada *stemming* Bahasa Indonesia.