

# Pengaruh *Text Preprocessing* terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)

Syifa Khairunnisa<sup>1</sup>, Prof. Dr. Adiwijaya, S.Si., M.Si.<sup>2</sup>, Said Al Faraby, S.T., M.Sc.<sup>3</sup>

<sup>1,2,3</sup>Fakultas Informatika, Universitas Telkom, Bandung

<sup>1</sup>syifakhnsa@students.telkomuniversity.ac.id, <sup>2</sup>adiwijaya@telkomuniversity.ac.id,

<sup>3</sup>saidalfaraby@telkomuniversity.ac.id

---

## Abstrak

COVID-19 merupakan pandemi yang meresahkan banyak masyarakat. Hal ini memunculkan banyak komentar masyarakat di sosial media Twitter. Komentar tersebut digunakan untuk analisis sentimen sehingga diketahui polaritas sentimen yang muncul, apakah positif, negatif, atau netral. Permasalahan ketika menggunakan data twitter ialah, data *tweet* masih mengandung banyak kata yang tidak baku seperti penulisan yang disingkat karena keterbatasan maksimal karakter yang dapat digunakan dalam sekali membagikan *tweet*. *Preprocessing* merupakan tahapan awal yang paling penting dalam analisis sentimen saat menggunakan data twitter, karena berpengaruh terhadap hasil performansi klasifikasi. Penelitian ini membahas secara khusus mengenai teknik *preproceesing* dengan melakukan beberapa skenario pengujian kombinasi teknik *preprocessing* untuk mengetahui teknik *preprocessing* yang menghasilkan akurasi paling optimal serta pengaruhnya terhadap analisis sentimen. Ekstraksi fitur menggunakan N-Gram dan pembobotan kata menggunakan TF-IDF. Mutual Information sebagai metode seleksi fitur. Metode klasifikasi yang digunakan adalah SVM karena mampu mengklasifikasikan data berdimensi tinggi sesuai dengan data yang digunakan pada penelitian ini yaitu berupa data teks. Hasil penelitian ini menunjukkan bahwa kinerja terbaik diperoleh dengan menggunakan kombinasi *cleaning* dan *stemming*; dan normalisasi kata, *cleaning*, dan *stemming* dengan akurasi yang sama sebesar 77.77%. penggunaan unigram menghasilkan akurasi yang lebih tinggi dibandingkan dengan bigram. Mutual Information mampu mengurangi masalah *overfitting* dengan berkurangnya fitur yang tidak relevan sehingga akurasi train dan test cukup stabil.

Kata kunci : COVID-19, twitter, analisis sentimen, *preprocessing*, *support vector machine*

---