

Abstract—Dalam studi ini, kami melakukan analisis sentimen terhadap tweet campuran bahasa Indonesia-Sunda. Bahasa Sunda adalah salah satu bahasa daerah Indonesia dengan lebih dari 42.000.000 penutur. Kami menggunakan model bahasa yang telah dibuat sebelumnya, IndoBERT, untuk menangani tugas analisis sentimen tersebut. Hasil evaluasi kami menunjukkan bahwa akurasi terbaiknya adalah 81%. Kami menganalisis kesalahan dan menemukan bahwa tweet yang salah diprediksi kebanyakan dikarenakan oleh kata-kata pada tweet dengan suatu label terdapat banyak di label lainnya. Mungkin juga karena kalimat dalam tweet terkesan ambigu, kata-kata yang digunakan dalam tweet tidak tersedia dalam kumpulan data latihan, atau penggunaan kata-kata yang disingkat dalam tweet.

Index Terms—sentiment analysis, IndoBERT, code-mixed data, Indonesian-Sundanese code mixed tweets, natural language processing