1. Pendahuluan

Twitter adalah salah satu layanan mikroblog terbesar yang memungkinkan pengguna dapat mengirim dan menerima sebuah *tweet*. Dengan *twitter* menjadi sebuah layanan untuk pengguna dapat menulis apapun, akan muncul berbagai topik didalamnya [1]. Oleh karena itu, *twitter* dapat dimanfaatkan untuk melakukan penelitian mengenai klasifikasi topik. Penelitian terkait klasifikasi topik sudah pernah dilakukan sebelumnya seperti penelitian dengan menggunakan *Word Embedding* pada klasifikasi topik [2].

Klasifikasi topik adalah menetapkan suatu topik kepada suatu data atau dokumen yang umumnya datanya berbentuk teks tidak terstrukur. Menetapkan suatu topik kepada suatu dokumen memungkinkan untuk melakukan pencarian, perhitungan statistik, dan pengklasifikasian yang bermakna [3]. Dalam penelitian ini, klasifikasi topik mengambil satu set dokumen teks (korpus *tweet*) sebagai input dan mengembalikan satu set topik ke korpus untuk menggambarkan topik apa yang ada didalam setiap *tweet* dalam korpus tersebut. Ada tantangan lain dalam melakukan klasifikasi topik pada *tweet*. *Tweet* hanya dibatasi 280 karakter didalamnya yang menyebabkan banyak tata Bahasa yang tidak benar, banyak kata gaul, dan *tweet* yang singkat [2]. Penggunaan kata gaul pada tweet dapat meningkatkan kemungkinan ketidakcocokan kosakata dan membuat *tweet* sulit dipahami topik apa yang ada didalamnya [4]. Maka dari itu, penelitian ini mengimplementasikan ekspansi fitur untuk mengatasi masalah dari faktor-faktor yang disebutkan sebelumnya.

Tujuan penelitian ini untuk melihat pengaruh ekspansi fitur dengan *fastText* pada kasus klasifikasi topik. Dalam rangka memenuhi tujuan ini, kami melakukan pengujian skenario *split data* pada dataset *tweet* dan menentukan pemilihan jumlah fitur yang terbaik pada TFIDF untuk dijadikan model *baseline*. Kemudian, setelah didapatkan model *baseline*, dilakukan pengujian ekspansi fitur dengan *fastText*. *FastText* dipenelitian ini menghitung *similarity* antar kata yang ada di korpus lalu dilakukan proses ekspansi fitur untuk setiap kata yang *similarity* kata nya ada pada dalam *tweet*, hal ini berguna untuk menambah informasi dan mengurangi ketidakcocokan kosakata pada *tweet* yang umumnya singkat (karakter yang terbatas), banyak kata gaul, dan tata Bahasa yang tidak benar. Kemudian, dilakukan *hyperparameter tuning* pada GBDT. Terakhir, *Gradient Boosted Decision Tree* dibandingkan dengan klasifikasi *ensemble learning* yang lain (*Ada Boost Decision Tree* dan *Decision Tree*).

Jurnal tugas akhir ini disusun sebagai berikut, bab 1 menjelaskan pendahuluan, kemudian bab 2 akan membahas teori/studi/literatur yang mendukung atau berkaitan erat dengan penelitian ini. Bab 3 akan membahas rancangan dan sistem yang dihasilkan. Bab 4 akan membahas hasil pengujian dan analisis hasil pengujian. Terakhir, bab 5 menjelaskan kesimpulan dan saran.