BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Hate speech atau ujaran kebencian adalah suatu bentuk ekspresi yang menyebar, menghasut, mempromosikan atau membenarkan kebencian, kekerasan dan diskriminasi terhadap seseorang atau sekelompok orang karena berbagai alasan (Davidson et al., 2017). Ujaran kebencian tersebut menjadi salah satu penyebab utama terjadinya gangguan pada kesehatan mental seseorang yang akhirnya dapat menyebabkan terjadinya depresi ataupun bunuh diri (Fauzi, 2019). Oleh sebab itu, ujaran kebencian juga menjadi salah satu penyebab tingginya angka prevalensi depresi di Indonesia yaitu sebesar 6,1 persen atau 11.315.500 orang untuk kelompok usia lebih dari 15 tahun (Ika, 2019). Salah satu wadah yang dijadikan tempat untuk menyebar ujaran kebencian adalah media sosial Twitter (Nursalikah, 2018).

Akan tetapi, mencegah penyebaran ujaran kebencian dalam media sosial Twitter bukan merupakan suatu hal yang mudah. Pendiri sekaligus CEO Twitter, Jack Dorsey, mengaku belum menemukan solusi terbaik untuk memecahkan masalah tersebut karena keterbatasan sumber daya untuk memeriksa teks dan konten dari setiap akun Twitter yang ada (Nistanto, 2018). Salah satu solusi yang dapat diambil adalah menyortir opini dalam komentar-komentar yang ada pada Twitter agar opini yang bersifat negatif dapat dihilangkan. Penentuan polaritas positif atau negatifnya suatu opini dapat dilakukan secara manual. Akan tetapi seiring bertambahnya sumber

opini, tentunya semakin banyak juga waktu dan usaha yang dibutuhkan untuk mengklasifikasikan polaritas opini tersebut. Oleh sebab itu, diajukan penerapan metode pembelajaran mesin untuk mengklasifikasi polaritas opini dari sumber data yang sangat banyak tersebut (Nurhuda et al., 2013). Salah satu metode pembelajaran mesin yang dapat digunakan untuk mengklasifikasikan polaritas dari suatu opini dalam jumlah besar secara otomatis adalah *Text Classification*.

Text Classification merupakan salah satu bidang dalam *Natural Language Processing* (NLP) yang mengotomasikan pengklasifikasian teks ke satu atau lebih kategori yang tepat berdasarkan isinya dengan membangun model menggunakan data latih (Mowafy et al., 2018). Salah satu metode *text classification* yang telah terbukti baik untuk teks Bahasa Indonesia adalah Support Vector Machine (SVM) (Rizaldy dan Santoso, 2017; Noviantho et al., 2017; Haryanto et al., 2018). Akan tetapi, SVM merupakan metode *supervised learning* yang memerlukan banyak data yang diklasifikasi secara manual. Hal tersebut tentunya membutuhkan banyak waktu dan tenaga dalam penerapannya yang menghilangkan tujuan dari penerapan pembelajaran mesin. Oleh sebab itu, metode *semi-supervised learning* menjadi solusi atas permasalahan tersebut, dimana banyak data yang perlu diklasifikasi secara manual jauh lebih sedikit. Salah satu metode *semi-supervised learning* yang terbukti baik untuk *text classification* adalah Bidirectional Encoder Representations from Transformers (BERT).

Beberapa penelitian sebelumnya (Kumar dan Ojha, 2019; Melamud et al., 2019; González-Carvajal dan Garrido-Merchán, 2020) menunjukkan bahwa BERT memiliki performa model yang lebih baik dari SVM. Metode Unsupervised Data

Augmentation (UDA) merupakan pengembangan dari BERT yang membutuhkan 1.250 kali lebih sedikit data teks yang telah diklasifikasi secara manual jika dibandingkan dengan BERT (Xie et al., 2019). Sehingga menggunakan UDA, waktu dan tenaga yang dibutuhkan untuk melakukan pemrosesan data teks juga jauh lebih sedikit.

Berdasarkan masalah dan penelitian yang telah disebutkan sebelumnya, maka penelitian ini akan mengimplementasikan metode Unsupervised Data Augmentation dengan melakukan tahap-tahap *text preprocessing* terlebih dahulu dan mengukur performa model dalam mendeteksi teks *hate speech* pada media sosial Twitter dengan metrik evaluasi.

1.2 Rumusan Masalah

Berdasarkan pemaparan dari latar belakang masalah, rumusan masalah dalam penelitian "Implementasi Metode Unsupervised Data Augmentation untuk Deteksi Teks Hate Speech dalam Bahasa Indonesia" adalah sebagai berikut.

1. Bagaimana cara mengimplementasikan metode Unsupervised Data Augmentation untuk mengklasifikasikan suatu komentar Twitter termasuk ke dalam kategori *hate speech* atau tidak?
2. Bagaimana performa metode Unsupervised Data Augmentation dalam mengklasifikasikan teks *hate speech* pada media sosial Twitter?

1.3 Batasan Masalah

Terdapat beberapa batasan masalah dalam penelitian yang dilakukan. Adapun batasan-batasan masalah dari penelitian adalah sebagai berikut.

1. Performa metode Unsupervised Data Augmentation yang dihasilkan dalam penelitian ini merupakan hasil dari pelatihan dan pengujian model dengan menggunakan data komentar Twitter yang didapat melalui penelitian Ibrohim dan Budi (2019).
2. Performa metode Unsupervised Data Augmentation yang dihasilkan dalam penelitian ini menggunakan kode program Unsupervised Data Augmentation yang didapat melalui *Github Repository* yang dibuat oleh Yun (2019).
3. *Pretrained* BERT model yang digunakan dalam penelitian ini menggunakan BERT-Base model yang didapat melalui penelitian yang dilakukan oleh Willie et al. (2020).

1.4 Tujuan Penelitian

Tujuan dari penelitian "Implementasi Metode Unsupervised Data Augmentation untuk Deteksi Teks Hate Speech dalam Bahasa Indonesia" adalah sebagai berikut.

1. Mengimplementasikan metode Unsupervised Data Augmentation untuk mengklasifikasikan suatu komentar Twitter termasuk ke dalam kategori *hate speech* atau tidak.

2. Mengukur performa metode Unsupervised Data Augmentation dalam mengklasifikasikan teks *hate speech* pada media sosial Twitter dengan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1 score*.

1.5 Manfaat Penelitian

Manfaat dari penelitian "Implementasi Metode Unsupervised Data Augmentation untuk Deteksi Teks Hate Speech dalam Bahasa Indonesia" adalah sebagai berikut.

1. Membangun sistem deteksi *hate speech* yang dapat membantu mengurangi komentar-komentar negatif dalam media sosial Twitter dengan melakukan penyortiran komentar berdasarkan polaritas opini.
2. Membantu pemerintah untuk mengurangi angka prevalensi depresi dan bunuh diri di Indonesia dengan menggunakan sistem yang dibangun untuk mendeteksi *hate speech* dalam media sosial Twitter yang merupakan salah satu penyebab utama seseorang mengalami depresi.
3. Membangun sistem klasifikasi *hate speech* dengan meminimalisir banyaknya penggunaan *supervised data*, sehingga mengurangi *effort* yang dibutuhkan untuk mengklasifikasi label secara manual dalam pembangunan sistem klasifikasi.