BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Di era modern ini internet dapat dikatakan telah mempengaruhi seluruh dunia dengan kehadirannya. Kecepatan dan daya sebarnya yang sangat luas menjadikan internet sebagai media penyebaran informasi yang sangat digemari oleh semua pihak [1]. Internet membuat sebuah tren baru dikalangan penggunanya yang disebut "*meme". Meme* dapat dikatakan sebagai replikasi gambar dengan sebuah teks didalamnnya dan bersifat kontekstual dengan topik yang sedang dibahas. Seiring dengan perkembangan zaman *meme* pun berkembang tidak hanya sebatas gambar yang digunakan untuk lelucon semata tetapi *meme* dapat berisikan pesan tentang menyinggung keadaan sosial masyarakat maupun mengandung muatan politik didalamnya yang biasa digunakan untuk mengomentari para petinggi negara dan kebijakannya [1]. Contoh dari *meme* dapat dilihat pada Gambar 1.1. Contoh memes yang diambil dari *hateful memes dataset* Dalam beberapa kasus bahkan penggunaan *meme* menyebabkan munculnya kasus *hate speech* seperti pada kasus *meme* yang menyinggung polisi di daerah ponorogo pada tahun tanggal 6 november 2015 yang lalu [2].

Gambar 1. 1 Contoh *memes* yang diambil dari *hateful memes dataset*

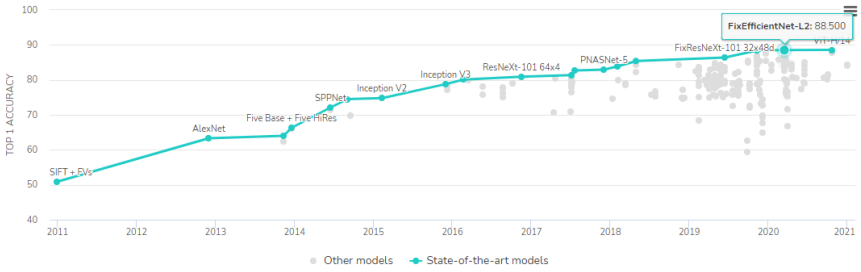
Terdapat begitu banyak kasus *hate speech* yang terjadi di Indonesia, ada sekitar 3.325 kasus yang ditangani oleh POLRI Selama tahun 2017 atau dapat dikatakan angka kasus tersebut naik sebanyak 44,99% dari tahun sebelumnya yang hanya terdapat 1.829 kasus *hate speech* [3]. Dari fenomena terjadinya *memes* yang mengandung *hate speech* terdapat beberapa dampak negatif yang ditimbulkan seperti, merusak hubungan antar manusia dan juga antar umat beragama [4], dan juga timbulnya fenomena *haters* [5].

Dengan banyaknya kasus *hate speech* yang terjadi di Indonesia dan populernya tren *meme* di Indonesia, penelitian ini akan membuat sebuah *Image Classification* sehingga bisa membedakan *meme* yang berpotensi untuk menyebabkan *hate speech* dan *meme* yang tidak berpotensi, menggunakan

FixEfficientNet-L2. *Image Classification* adalah suatu tugas dalam bidang *computer vision* dengan tujuan untuk memasukkan gambar kedalam klasifikasi berdasarkan konten visual dari gambar tersebut [6].

Deep learning adalah sebuah bentuk dari *machine learning* yang memungkinkan komputer untuk belajar dari pengalaman dan mengerti dunia dengan konteks sebuah konsep hirarki, dan salah satu dari banyak bagian *deep learning* adalah *Convolutional Neural Network* [7]. *Convolutional Neural Network* dibuat untuk memproses data yang berbentuk *multiple array* seperti sebuah gambar berwarna yang terdiri dari tiga *array* 2D bermuatan intensitas *pixel* dalam 3 *channel* warna. Terdapat banyak data yang berbentuk *multiple array* seperti 1D untuk sinyal dan *sequences*, termasuk Bahasa; 2D untuk gambar atau *spectrograms* dari audio; dan 3D untuk video atau gambar *volumetric*. Sejak tahun 2000an *Convolutional Neural Network* telah mendapatkan kesuksesan besar dalam hal mendeteksi, segmentasi dan pengenalan sebuah object atau wilayah pada gambar-gambar, tugas tersebut dilakukan dengan data yang berlabel relatif banyak [8]. *Convolutional Neural Network* memiliki neuron tiga dimensi yaitu lebar, tinggi, dan kedalaman. Lebar dan tinggi disebut juga sebagai ukuran lapisan, dan kedalaman disebut juga sebagai jumlah lapisan. *Convolutional Neural Network* dapat terdiri dari puluhan, ratusan, bahkan ribuan lapisan yang berfungsi untuk mempelajari pola-pola yang ada pada gambar sehingga dapat mengidentifikasi gambar tersebut. Secara umum *Convolutional Neural Network* memiliki dua lapisan yang terdiri dari lapisan pertama atau disebut juga *feature extraction layer* dan

lapisan kedua atau disebut juga *classification layer* [9]. Arsitektur dari *convolutional neural network* terdapat banyak jenis, salah satunya adalah arsitektur yang akan digunakan pada penelitian ini yaitu *FixEfficientNet-L2*. Alasan utama dipilihnya *FixEfficientNet-L2* pada penelitian ini dikarenakan *FixEfficientNet-L2* menjadi *state-of-the-art* dalam *ImageNet challenge* pada pertengahan tahun 2020 dengan akurasi sebesar 88.5% [10].



Gambar 1. 2 Grafik *state-of-the-art model* dalam *ImageNet challenge* [11]

Penelitian kali ini dilakukan dengan menggunakan data *Hateful Memes* dari facebook yang memang bertujuan untuk membuat tantangan untuk para peneliti dalam mendeteksi *meme* yang berpotensi untuk menjadi *hateful meme*, dan pembuatan dari penelitian ini juga berdasarkan salah satu penelitian yang berjudul "*The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*" yang dibuat oleh tim *Facebook AI*.

1.2 Batasan Masalah

Berikut adalah beberapa batasan masalah pada penelitian ini:

1. *Model* yang dibangun hanya dilatih untuk mengidentifikasi *memes* yang menggunakan Bahasa Inggris.
2. *Model* dilatih dengan 8.500 dataset.
3. *Model* mempunyai dua kelas klasifikasi yaitu, "*hateful*" dan "*non-hateful*".
4. Metode yang digunakan adalah *convolutional neural network* dengan *architecture FixEfficientNet-L2.*

1.3 Rumusan Masalah

Rumusan masalah pada penelitian kali ini yang disimpulkan berdasarkan latar belakang masalah adalah sebagai berikut:

1. Bagaimana kinerja dari *model image classification* yang dibuat dengan menggunakan *FixEfficientNet-L2* terhadap klasifikasi *hateful memes dataset* dibandingkan dengan penelitian terdahulu?

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Tujuan dari dilakukannya penelitian ini adalah sebagai berikut:

1. Membuat sebuah *model* yang dapat mengidentifikasi *meme* berupa gambar yang mengandung *hate speech* menggunakan *FixEfficientNet-L2.*
2. Mengetahui kinerja *FixEfficientNet-L2* dalam mengidentifikasi *meme* yang berpotensi mengandung *hate speech*

1.4.1 Manfaat Penelitian

Berikut adalah beberapa manfaat dari penelitian ini:

1. Menyumbangkan gagasan *model* untuk penelitian selanjutnya dalam topik *hate speech detection*.
2. Dapat menjadi referensi untuk penindak tindakan *hate speech* di media sosial.