BAB I
PENDAHULUAN

1.1 Latar Belakang

Kehidupan manusia sehari-hari saat ini tidak terlepas dari peran media massa sebagai alat komunikasi yang menyampaikan informasi secara cepat dan luas. Salah satu media yang popular yaitu youtube. Saat ini, Youtube menjadi media komunikasi dan sarana informasi modern. Youtube didirikan pada tahun 2005 oleh Steve Chen, Chad Hurley, dan Jawed Karim. Youtube kemudian diakuisisi oleh Google Inc. pada November 2006, dan youtube kini dimiliki oleh Google Inc [1]. Youtube menyediakan berbagai konten menarik bagi penggunanya, termasuk *podcast* yang menawarkan ragam topik dan pembicaraan yang menginspirasi dan menghibur.

Podcast merupakan salah satu media massa yang populer di Indonesia. Indonesia menempati posisi kedua sebagai negara dengan jumlah pendengar *podcast* terbanyak di dunia[2]. *Podcast* adalah sarana komunikasi massa. *Podcast* dengan program dan kontennya yang semakin beragam dapat memberikan dampak yang signifikan dalam mengubah kehidupan masyarakat, baik secara langsung maupun tidak langsung. Keberhasilan sebuah *podcast* dalam menarik perhatian pemirsa dan pendengar tidak terlepas dari keberhasilan program dan konten acara yang ada di dalam *podcast* tersebut [3].

Podcast digunakan untuk memberikan hiburan, menemani pendengar melakukan aktivitas lain, atau sekadar mengisi waktu luang yang membosankan. *Podcast* memberi pendengar konten menarik yang tak ada habisnya. *Channel podcast* populer Indonesia ini memiliki beragam topik konten mulai dari komedi, inspirasi, bisnis, olahraga, horor, dan topik acak lainnya yang dibahas di setiap saluran. Namun, sebagian besar *podcast* ini adalah percakapan antara dua orang atau lebih yang membahas topik tertentu. Percakapan antara dua orang atau lebih memberikan informasi kepada pendengarnya. Informasi yang ditampilkan mungkin berkaitan dengan topik, pendidikan, bahkan hiburan [4].

Percakapan antara dua orang atau lebih dalam membahasa topik tertentu tidak lepas dari penggunaan kata atau kalimat yang bebas. Dalam berkomunikasi, kesantunan berkata-kata perlu kita perhatikan, karena yang kita perhatikan bukan

hanya pada aspek pengertian saja, namun juga pada aspek keselarasan antara penutur dan petutur [5]. Kebebasan berbicara dalam konten *podcast* tidak lepas dari bahasa yang kasar. Terlepas dari kebebasan berekspresi, konten yang penuh bahasa kasar cenderung meningkatkan ketegangan dan ketidaknyamanan di antara pendengar [6]. Ini bisa mengurangi kualitas pengalaman mendengarkan dan mengganggu kesan positif yang ingin dicapai oleh para pembuat konten. Selain itu, bahasa yang kasar atau mengandung kebencian dapat memicu reaksi negatif dari masyarakat, memperburuk masalah sosial, dan menyebarkan budaya tidak sehat.

Untuk menciptakan lingkungan internet yang sehat, terutama di youtube, diperlukan sistem yang mampu mendeteksi bahasa kasar khususnya dalam konten *podcast*. Oleh karena itu, pengembangan teknologi semacam ini sangat dibutuhkan agar konten tetap berkualitas dan bebas dari ujaran yang tidak pantas. Salah satu pendekatan yang dapat digunakan adalah dengan memanfaatkan algoritma BERT (Bidirectional Encoder Representations from Transformers) dalam pemrosesan bahasa alami (Natural Language Processing/NLP). *Bidirectional Encoder Representations from Transformers* atau biasa disebut BERT merupakan algoritma deep learning yang dirilis oleh Google yang masih relevan dengan NLP (Natural Language Processing). Algoritma ini merupakan intrusi ke dalam model Transformer, yang memproses kata-kata dalam sebuah kalimat berdasarkan apakah ada hubungan antara kata tersebut dan keseluruhan kalimat. Cara kerja algoritma BERT berbeda dengan algoritma pemrosesan bahasa lainnya. BERT mengolah kata dengan mempelajari konteks kata dari kata yang ada [7]. Algoritma BERT menangani semua konteks dengan melihat pola yang muncul sebelum dan sesudah kata.

Penelitian terdahulu menunjukkan potensi BERT dalam berbagai jenis deteksi berupa teks. Imroatus, misalnya, memanfaatkan BERT untuk mendeteksi berita hoaks, yang digabungkan dengan Random Forest Classifier guna mengatasi masalah overfitting, dan menghasilkan akurasi sebesar 67% dalam klasifikasi pesan disinformasi[8]. Penelitian lain menggunakan BERT dalam mendeteksi penghinaan berbasis orientasi seksual atau kepribadian, misalnya untuk komentar homofobik dalam bahasa Turki. Sistem ini mampu mencapai F1-score sebesar

82,64% dalam mendeteksi homofobik dan 91,75% dalam mendeteksi kebencian. Studi-studi ini menunjukkan bahwa algoritma BERT mampu memahami konteks kompleks dan nuansa bahasa, menjadikannya relevan untuk tugas seperti deteksi bahasa kasar dalam podcast[9].

Dikarenakan konten *podcast* digemari oleh semua kalangan masyarakat dan semua usia, serta menyediakan berbagai jenis topik seperti hiburan, politik dan lainnya. Sehingga konten *podcast* idealnya bermuatan postif, tetapi konten *podcast* banyak yang menggunakan bahasa kasar. Berdasarkan latar belakang yang telah dijelaskan, penelitian ini bertujuan untuk mendeteksi bahasa kasar yang dapat terkandung dalam konten *podcast*. Dengan judul penelitian ”Deteksi bahasa kasar pada konten *podcast* youtube menggunakan algoritma *bidirectional encoder representations from transformers* (BERT)” diharapkan dapat memberikan edukasi dan informasi kepada masyarakat, sehingga dapat memilah dan memilih konten *podcast* yang tidak mengandung berbahasa kasar.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang dijabarkan di atas, penulis memiliki beberapa rumusan masalah terkait latar belakang tersebut, yaitu :

1. Bagaimana implementasi algoritma BERT untuk mendeteksi bahasa kasar pada konten *podcast* youtube?
2. Bagaimana persentase kemunculan bahasa kasar pada konten *podcast* youtube yang telah dideteksi?

1.3 Batasan Masalah

Berdasarkan rumusan masalah yang telah dijabarkan di atas, penulis membatasi masalah yang akan dianalisa pada pembuatan system ini. Batasan-batasan tersebut yaitu :

1. Konten youtube yang dideteksi berasal dari *podcast.*
2. Konten *podcast* yang membahas tentang esport, membership youtube, timnas Indonesia dan Kesehatan.
3. Konten *podcast* berbahasa Indonesia.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang dijabarkan di atas, penulis memiliki beberapa tujuan terkait latar belakang tersebut, yaitu:

1. Merancang dan mengimplementasikan algoritma BERT untuk mendeteksi bahasa kasar pada konten *podcast*.
2. Mengetahui persentase kemunculan bahasa kasar pada konten *podcast*.

1.5 Manfaat Penelitian

Penulis berharap penelitian ini dapat memberikan manfaat sebagai berikut:

1. Bagi Pembaca

Penelitian ini dapat memberikan manfaat bagi pembaca mengenai berapa persentase kemunculan bahasa kasar pada konten *podcast* youtube dan menilai bagaimana konten *podcast* yang selalu menggunakan bahasa kasar.

2. Bagi Penulis

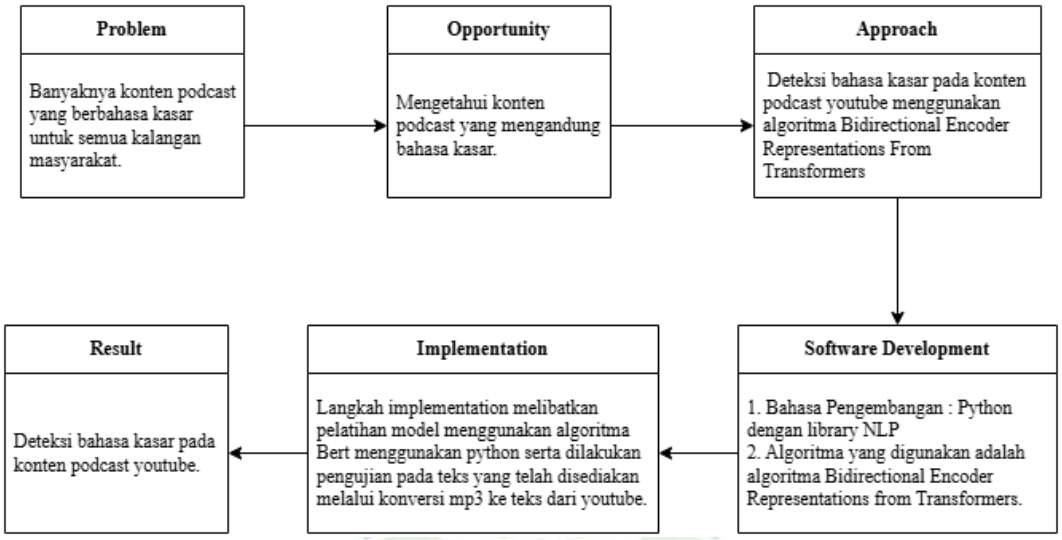
Sebagai kontribusi dalam pengembangan sistem menggunakan algoritma BERT untuk mengetahui berapa persentase kemunculan bahasa kasar pada konten *podcast* youtube.

3. Bagi Akademik

Penelitian ini dapat menjadi sumber referensi dan dasar penelitian lebih lanjut dalam bidang AI menggunakan algoritma BERT. Pengembangan lebih lanjut dalam bidang AI terutama menggunakan algoritma BERT untuk pemrosesan kalimat.

1.6 Kerangka Pemikiran

Adapun kerangka pemikiran dari penelitian disajikan pada Gambar 1.1.



Gambar 1. 1 Kerangka pemikiran

1.7 Metodolgi penelitian

1.7.1 Teknik Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini dimulai dengan menentukan tema terlebih dahulu. Terdapat 4 tema yang dipilih, yaitu *esport*, membership youtube, tim nasional Indonesia, dan kesehatan. Setelah dilakukan pemilihan tema, langkah selanjutnya adalah mengidentifikasi video di youtube yang relevan dengan masing-masing tema. Video-video tersebut kemudian diunduh untuk keperluan penelitian, namun sebelum melakukan pengunduhan dan analisis, peneliti menghubungi pemilik video melalui email yang tercantum di deskripsi video untuk meminta izin resmi menggunakan konten tersebut dalam penelitian ini. Proses pengiriman email mencakup penjelasan mengenai tujuan penelitian serta bagaimana data dari video tersebut akan digunakan, untuk memastikan kepatuhan terhadap etika penelitian dan hak kekayaan intelektual.

1.7.2 Model Pengembangan

Metode pengembangan yang digunakan dalam penelitian ini adalah kerangka kerja CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Metode ini sering digunakan dalam pengembangan proyek data mining dan analisis data. Kerangka kerja ini terdiri dari enam tahap utama yang memberikan panduan terstruktur dalam proyek. Tahapan tersebut diawali dengan pemahaman tentang konteks penelitian (*business understanding*), diikuti oleh pemahaman terhadap data yang akan digunakan (*data understanding*), persiapan data (*data preparation*), tahap pemodelan (*modelling*), evaluasi model (*evaluation*), dan diakhiri dengan penerapan hasil penelitian (*deployment*)[10].

1.8 Sistematika Penulisan

BAB I PENDAHULUAN

Pada bab ini peneliti menjelaskan latar belakang penelitian, rumusan masalah penelitian, Batasan masalah penelitian, manfaat penelitian serta kerangka penelitian yang akan dilakukan.

BAB II KAJIAN LITERATUR

Pada bab ini, peneliti menjelaskan penelitian terdahulu yang berkaitan dengan penggunaan algoritma BERT. Selain itu, peneliti juga memaparkan beberapa teori yang mendukung penelitian ini, seperti teori deep learning, teori BERT, penjelasan mengenai Python, teori tentang bahasa kasar, serta pembahasan mengenai SCHRIM-DM.

BAB III METODOLOGI PENELITIAN

Pada bab ini, peneliti memaparkan tahapan penelitian yang dilakukan, mulai dari pengumpulan data hingga pengembangan perangkat lunak. Peneliti juga menjelaskan langkah-langkah serta teknik yang digunakan dalam penelitian ini.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini, peneliti menyajikan hasil penelitian berdasarkan tahapan metode yang telah dilakukan. Pemaparan disusun sesuai dengan rumusan masalah

penelitian dan bertujuan untuk menjawab setiap permasalahan yang telah dirumuskan.

BAB V SIMPULAN DAN SARAN

Pada bab ini, peneliti menyimpulkan hasil pembahasan secara ringkas, dengan menjawab permasalahan yang telah dirumuskan sebelumnya. Selain itu, peneliti juga memberikan saran untuk penelitian lanjutan yang lebih mendalam.