BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi digital dan tingginya penggunaan media sosial seperti YouTube telah menciptakan ruang komunikasi interaktif yang terbuka bagi pengguna untuk saling berkomentar, berbagi, serta berinteraksi dalam skala besar. Namun, keterbukaan ini juga menghadirkan permasalahan baru berupa meningkatnya komentar spam yang mengandung tautan berbahaya, penipuan, hingga promosi aktivitas ilegal seperti judi online [1]. Fenomena komentar spam ini mengganggu kenyamanan pengguna dan merusak citra platform, sementara sistem moderasi bawaan YouTube belum sepenuhnya efektif dalam mendeteksi komentar berbahaya, terutama yang ditulis dalam bahasa Indonesia dengan karakter non-standar atau disamarkan [2].

Sejumlah penelitian telah dilakukan untuk menangani masalah komentar spam di media sosial. Abdurrohim dkk. [3] Melakukan implementasi algoritma *Support Vector Machine (SVM)* untuk klasifikasi komentar spam pada Instagram dan menemukan bahwa SVM dengan menggunakan kernel *RBF (Radial Basis Function)* dan parameter Gamma menghasilkan akurasi sebesar 96.82%. Penelitian lain oleh Tjahyadi dkk. [1] mengembangkan aplikasi deteksi komentar spam youtube berbasis web menggunakan algoritma SVM, yang menunjukkan akurasi sebesar 89,58%. Meskipun metode *machine learning* klasik tersebut cukup efektif, keduanya masih menghadapi keterbatasan dalam menangani komentar dengan gaya bahasa tidak baku, singkatan, serta penggunaan karakter atau simbol unik. Di sisi lain, Cahyo dkk. [2] memadukan metode *Regular Expression* dan *Fuzzy Matching* untuk mengenali variasi ejaan dan kesalahan penulisan pada komentar promosi judi online, mencapai akurasi hingga 90,85%.

Pendekatan yang lebih modern berbasis *deep learning* mulai dikembangkan untuk meningkatkan efektivitas klasifikasi teks. Samuel dan Kristiadi [4] menunjukkan bahwa model *Transformer* yang di-*fine-tune* dengan IndoBERT mampu mendeteksi teks promosi judi online dengan akurasi mencapai 97% dengan nilai precision dan recall yang seimbang. IndoBERT dipilih dan memiliki keunggulan utama dibandingkan model *Transformer* lain seperti *Multilingual* BERT (mBERT) karena dua alasan utama. Pertama, model ini mewarisi arsitektur dasar dengan mekanisme *Bidirectional Attention*, yang memungkinkannya membaca dan memahami konteks kalimat secara utuh dari dua arah sekaligus (depan ke

belakang dan sebaliknya). Kedua, model ini dirancang dan dilatih secara khusus menggunakan *corpus* besar berbahasa Indonesia, mencakup berbagai sumber teks seperti berita, media sosial, dan publikasi ilmiah [5]. Proses *pre-training* yang berfokus pada struktur, morfologi, dan sintaksis Bahasa Indonesia yang dipadukan dengan mekanisme *Bidirectional Attention* tersebut menjadikan IndoBERT lebih sensitif terhadap konteks lokal serta lebih adaptif dalam menangani bahasa tidak baku, singkatan, campuran bahasa (seperti "togel", "slot gacor", "spin gratis"), dan gaya penulisan khas media sosial yang sering digunakan dalam promosi judi online [4].

Meskipun penelitian terdahulu telah menunjukkan hasil akurasi tinggi, sebagian besar masih terbatas pada tahap pengujian model tanpa implementasi dalam bentuk sistem yang siap digunakan. Riza dkk. [6] mengimplementasikan algoritma Naïve Bayes untuk filtrasi komentar spam judi online di YouTube dan menghasilkan akurasi hingga 97,1%, namun penerapannya masih berbasis *Command Line Interface* (CLI) yang kurang praktis bagi pengguna umum. Keterbatasan tersebut menunjukkan perlunya pengembangan penelitian ke arah implementasi yang lebih aplikatif dan mudah diakses.

Berdasarkan uraian sebelumnya, maka pada tugas akhir ini diangkat topik penelitian berjudul "**IMPLEMENTASI APLIKASI DETEKSI KOMENTAR SPAM IKLAN JUDI ONLINE PADA YOUTUBE BERBAHASA INDONESIA MENGGUNAKAN INDOBERT**". Diharapkan dapat menjembatani kesenjangan antara hasil penelitian akademik dan kebutuhan nyata akan sistem deteksi komentar spam berbahasa Indonesia yang efektif dan mudah digunakan.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah bagaimana mengimplementasikan model IndoBERT untuk mendeteksi dan filtrasi komentar spam iklan judi online berbahasa Indonesia pada platform YouTube secara efektif melalui pengembangan aplikasi yang mudah diakses dan digunakan, serta bagaimana penerapannya dalam bentuk *chrome extension* dapat memberikan solusi praktis bagi pengguna dalam mendeteksi dan filtrasi komentar spam iklan judi online secara otomatis.

1.3 Tujuan

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan model IndoBERT dalam sistem deteksi komentar spam iklan judi online berbahasa Indonesia pada platform YouTube.

2. Mengembangkan aplikasi deteksi komentar spam dalam bentuk *chrome extension* yang praktis dan mudah digunakan oleh pengguna.
3. Menguji kinerja model IndoBERT dalam mengenali komentar dengan bahasa tidak baku, singkatan, serta variasi ejaan khas media sosial.

1.4 Manfaat

Hasil penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Bagi pengguna YouTube, aplikasi ini dapat membantu menciptakan lingkungan komentar yang lebih bersih dari spam iklan judi online.
2. Bagi akademisi dan peneliti, penelitian ini dapat menjadi kontribusi terhadap pengembangan model NLP berbahasa Indonesia dan penerapan IndoBERT dalam konteks keamanan digital.
3. Bagi masyarakat umum, hasil penelitian ini berkontribusi pada peningkatan literasi digital dan keamanan informasi di ruang publik daring.

1.5 Ruang Lingkup

Untuk menjaga fokus penelitian agar tetap terarah, ruang lingkup penelitian ini meliputi:

1. Dataset yang digunakan merupakan hasil dari proses *crawling* data komentar *channel* YouTube berbahasa indonesia yang dikumpulkan dari tanggal 15 oktober 2025 hingga 9 desember 2025. Proses pengambilan data dilakukan dalam dua tahap menggunakan YouTube Data API v3. Tahap pertama (Dataset Awal) mengumpulkan 3.152 komentar , dan Tahap Kedua (Dataset Utama) mengumpulkan 114.482 komentar, sehingga total *dataset* berjumlah 117.634 data.
2. Proses pelabelan data menggunakan strategi semi-supervised dua tahap. Tahap pertama adalah pelabelan manual pada 3.152 data awal. Tahap kedua adalah pelabelan otomatis (pseudo-labeling) pada 114.482 data baru menggunakan model awal (Model A) yang dilatih khusus pada data *undersampled* seimbang. Kategori label final adalah spam (iklan judi) dan non-spam
3. Model yang digunakan adalah IndoBERT (model Indobenchmark/Indobert-base-p1) yang telah dilatih khusus pada korpus bahasa Indonesia. Model ini di-*fine-tune* untuk tugas klasifikasi teks biner.
4. Evaluasi kinerja model menggunakan *confusion matrix* serta metrik *accuracy*, *precision*, *recall*, dan *F1-score*.

5. Implementasi akhir penelitian ini adalah sebuah Chrome Extension. Aplikasi ini dirancang dengan arsitektur *client-server*, menggunakan *frontend* (HTML/CSS/JavaScript) yang memanggil *backend* API (dibangun dengan Python Flask) untuk mengintegrasikan model IndoBERT yang telah dilatih dan melakukan deteksi *real-time* di halaman YouTube