BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Berkembangnya sektor industri *game* di Indonesia, menyebabkan sejumlah studio pengembang *video game* mulai bermunculan dan mulai mengembangkan berbagai jenis *video game*. Namun, dalam mengembangkan suatu video *game*, pengembang juga harus memperhatikan pendapat dari pengguna atau pemainnya agar dapat mempertahankan minat pengguna dalam memainkan video *game* tersebut. Pemahaman terhadap kebutuhan dan keinginan pengguna sangat penting, karena dengan memahami hal tersebut pengembang *video game* dapat merancang suatu *game* yang lebih efektif dan sesuai dengan keinginan pemainnya ketika memainkan *game* tersebut (Tan dkk., 2021). Salah satu teknik yang dapat dilakukan dalam memperhatikan kebutuhan dan keinginan pengguna tersebut adalah dengan memperhatikan ulasan terkait dengan video game sejenis yang telah dirilis sebelumnya.

Ulasan memiliki peranan yang sangat penting sebagai sumber dalam menyampaikan pendapat ataupun sentimen melalui pengguna dalam topik tertentu. Dalam konteks ini yakni terkait dengan *video game*, ulasan tersebut dapat dilihat melalui berbagai platform sosial media atau platform distribusi game. Salah satu media atau platform yang dapat menampilkan ulasan pemain terhadap suatu *video game* adalah *Steam*. Pada platform *Steam*, dalam memberikan ulasan pengguna *Steam* juga dapat memberikan peringkat dalam bentuk *"Recommended"* (positif) atau *"Not Recommended"* (negatif) (Guzsvinecz & Szűcs, 2023). Pada platform *Steam* ulasan pengguna dapat diakses secara publik, dan *Valve Corporation* selaku pengembang *Steam* juga menyediakan *Application Programming Interface* (API) untuk mendapatkan ulasan secara masif dan *open source* (Valve Corporation, 2022a).

Salah satu metode yang digunakan dalam mengekstraksi konten dalam suatu ulasan adalah dengan melakukan analisis sentimen, yang merupakan proses menganalisis teks dan mengidentifikasi sentimen secara otomatis menggunakan

kecerdasan buatan (AI) (Britto & Pacífico, 2020). Analisis sentimen dapat membantu pengembang *video game* untuk mengidentifikasi serta memberikan pandangan pemain terhadap video game mereka dalam bentuk positif, netral, atau negatif, yang kemudian dapat dirangkum dalam suatu dokumen terperinci (Kusnadi dkk., 2021). Sehingga dengan mengetahui sentimen tersebut pengembang *game* akan dapat memutuskan tindakan yang tepat ketika melakukan peningkatan ataupun pengembangan video *game* selanjutnya (Kusnadi dkk., 2021).

Analisis sentimen memiliki berbagai jenis model dan algoritma satunya adalah dengan menggunakan IndoBERT. IndoBERT merupakan salah model transformer yang dikembangkan dari model BERT *(Bidirectional Encoder Representations)*, yang merupakan model pembelajaran mendalam yang diciptakan menangani berbagai tugas pemrosesan bahasa alami (NLP) yang telah dilatih pada teks tanpa label, dan kemudian disesuaikan khusus untuk setiap tugas tertentu dengan memanfaatkan data berlabel (Kurniawan dkk., 2022; Paul & Saha, 2022). BERT juga telah dikembangkan dalam berbagai bahasa untuk dapat membantu pemrosesan bahasa alami terutama pada berbagai bahasa yang memiliki struktur kalimat berbeda. Salah satunya model IndoBERT yang digunakan untuk pemrosesan bahasa alami dalam suatu data berbahasa Indonesia.

IndoBERT memiliki arsitektur yang mirip dengan BERT, perbedaannya hanya terletak pada dataset yang digunakan untuk pelatihan *unsupervised* (Dharmawan dkk., 2023). IndoBERT yang dikembangkan oleh IndoNLU menjadi satu tren baru dalam pemrosesan bahasa alami (NLP) yang dikumpulkan dan dilatih menggunakan 4 miliar kata dalam Bahasa Indonesia, melalui berbagai sumber media online berbahasa Indonesia (Isa dkk., 2022). IndoBERT dibangun dengan kosa kata Bahasa Indonesia dan telah di *pre-training* menggunakan dataset Indo4B sehingga IndoBERT memiliki kinerja yang baik dalam memproses bahasa alami terutama yang menggunakan Bahasa Indonesia (Juarto & Yulianto, 2023). Setelah di *pre-training* IndoBERT juga dapat disesuaikan dan dilatih kembali menggunakan dataset lain yang memiliki label khusus sehingga dapat menghasilkan performa yang lebih baik, fase ini disebut dengan *fine-tuning* IndoBERT (Jayadianti dkk., 2022).

Model IndoBERT memiliki performa yang cukup baik dalam pemrosesan bahasa alami seperti pada penelitian yang dilakukan oleh Mahfudiyah dan Alamsyah pada tahun 2022 mengenai implementasi model IndoBERT terhadap ulasan layanan Gojek menggunakan model IndoBERT yang mendapatkan hasil akurasi hingga 96% (Mahfudiyah & Alamsyah, 2022). Namun, performa model IndoBERT juga sangat dipengaruhi oleh tingkat kompleksitas dataset yang digunakan, terutama dalam konteks data yang telah mengalami ekstraksi fitur dengan metode yang berbeda. Kompleksitas data dalam dataset dapat dilihat dari tingkat variasi dan keragaman informasi dalam dataset seperti jumlah variabel yang tinggi, nilai yang beragam, dan hubungan nonlinear.

Dalam melakukan analisis sentimen, ekstraksi fitur juga menjadi salah satu langkah yang penting yang dapat mempengaruhi performa suatu model atau algoritma, dimana ektraksi ini melibatkan pemilihan fitur untuk mengubah teks menjadi data yang dapat di klasifikasikan (Rusli dkk., 2020). Beberapa metode ektraksi fitur yang paling umum digunakan yakni *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Word2Vec*. TF-IDF merupakan salah satu algoritma dalam ektraksi fitur yang paling umum digunakan, dimana TF-IDF berfungsi sebagai penghitung bobot setiap kata yang umum digunakan (Pratomo dkk., 2021). Sedangkan *Word2Vec* merupakan sebuah algoritma NLP yang menggunakan model *neural network* untuk memahami asosiasi antara kata-kata dari korpus teks yang luas (Faisal dkk., 2021).

TF-IDF dan Word2Vec sudah cukup sering digunakan dalam proses ektraksi fitur yang sering dikomparasikan satu sama lain dan dikombinasikan dengan berbagai model atau algoritma yang berbeda. TF-IDF yang berfokus pada frekuensi kata dalam dokumen, sedangkan Word2Vec mempresentasikan kata sebagai vektor berdimensi tinggi, menangkap hubungan semantik dan konteks kata memiliki poin kelebihan dan kekurangan masing-masing. Dalam beberapa kasus, TF-IDF dapat menghasilkan hasil yang lebih baik daripada Word2Vec, dan begitu juga sebaliknya tergantung dengan model klasifikasi dan data yang digunakan.

Sebagai contoh penelitian terkait dengan perbandingan ekstraksi fitur TF-IDF dan Word2Vec sudah pernah dilakukan sebelumnya. Diantaranya adalah penelitian yang dilakukan oleh Cahyani dan Patasik di tahun 2021 mengenai

perbandingan performa antara TF-IDF dan Word2Vec dalam mengkalisfikasikan teks emosi menggunakan model SVM dimana didapatkan bahwa TF-IDF memiliki performa yang lebih baik dengan akurasi rata-rata 90% dibanding dengan Word2Vec yang memiliki akurasi rata-rata 88% (Cahyani & Patasik, 2021). Sedangkan penelitian lainnya yang dilakukan oleh Dewi dkk, pada tahun 2022 terkait analisis sentimen vaksinasi COVID-19 yang didapat melalui twitter dengan membandingkan metode TF-IDF dan *Word2Vec* menggunakn model RNN, didapatkan bahwa *Word2Vec* memberikan tingkat akurasi yang lebih tinggi dibanding TF-IDF dengan nilai pada Word2Vec yakni 53% sedangkan TF-IDF 51% (Dewi dkk., 2022).

Oleh karena itu, diperlukan perbandingan antara metode ekstraksi fitur TF-IDF dan Word2Vec menjadi suatu kebutuhan yang esensial ketika diterapkan pada model klasifikasi teks yang lebih mutakhir, yakni IndoBERT. Hal ini bertujuan untuk menilai baik antara performa IndoBERT dalam memproses dataset yang telah mengalami ekstraksi fitur dengan metode TF-IDF dan Word2Vec. Maupun untuk mengetahui performa dari masing-masing ektstraksi fitur tersebut ketika melakukan klasifikasi menggunakan model IndoBERT, sehingga membantu dalam memahami kelebihan dan kekurangan masing-masing metode dan membantu dalam memilih metode yang paling sesuai pada kasus tertentu.

Berdasarkan uraian latar belakang yang telah dipaparkan tersebut, maka pada penelitian ini akan dirancang dengan tujuan membandingkan metode ekstraksi fitur TF-IDF dan *Word2Vec* menggunakan model IndoBERT untuk mengetahui performa masing-masing metode ekstraksi tersebut ketika diproses dengan menggunakan model indoBERT. Selain penelitian ini akan menguji performa model indoBERT ketika dikombinasikan dengan metode ektraksi fitur yang. Oleh karena itu dirancang sebuah penelitian dengan judul Analisis Sentimen Perbandingan Ektraksi Fitur TF-IDF dan Word2Vec Pada Analisis Sentimen Menggunakan Model *Fine-Tuning* IndoBert untuk Ulasan Game Lokal di Steam yang diharapkan dapat memberikan hasil sentimen terkait dengan ulasan *game* lokal di *Steam* menggunakan model indoBERT dengan performa yang lebih optimal.

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang masalah yang dipaparkan sebelumnya, maka rumusan masalah dari penelitian ini didapatkan sebagai berikut :

1. Bagaimana perbandingan performa hasil metode TF-IDF dan Word2Vec dalam mengklasifikasikan sentimen pengguna terkait ulasan *game* lokal pada *platform Steam* menggunakan model IndoBERT?
2. Bagaimana hasil analisis sentimen data ulasan pengguna terkait *game* lokal yang ada *platform Steam* menggunakan model IndoBERT?

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah yang diuraikan sebelumnya, maka tujuan dari penelitian ini dapat dijabarkan sebagai berikut :

1. Untuk mengetahui perbandingan performa hasil metode TF-IDF dan Word2Vec dalam mengklasifikasikan sentimen pengguna terhadap ulasan *game* lokal pada *platform Steam* menggunakan model indoBERT.
2. Untuk mengetahui hasil analisis sentimen data ulasan pengguna terkait *game* lokal yang ada *platform Steam* menggunakan model indoBERT.

1.4. Batasan Masalah Penelitian

Untuk menjadikan penelitian lebih terfokus, berikut adalah batasan-batasan masalah yang ditetapkan dalam penelitian ini:

1. Kumpulan data *(dataset)* yang digunakan dalam penelitian ini bersumber dari kumpulan ulasan mengenai video game buatan *developer* Indonesia yang ada pada *platform Steam Store*.
2. Data ulasan yang digunakan adalah ulasan berbahasa Indonesia yang terdapat pada video game buatan *developer* Indonesia yang dirilis dari tanggal 1 januari 2020 hingga 31 juli 2023.

1.5. Manfaat Penelitian

Berdasarkan permasalahan yang telah diuraikan dan tujuan penelitian yang telah ditetapkan, penelitian ini diharapkan dapat memberikan manfaat berikut:

1. Manfaat bagi Pembaca:
    a. Penelitian ini dapat memberikan informasi terkait ulasan pengguna dalam merespons dan mengevaluasi *game* lokal yang tersedia di *platform Steam*.
    b. Penelitian ini dapat menjadi bahan evaluasi yang berguna bagi pengembang *game* asal Indonesia agar dapat mengembangkan *game* yang lebih sesuai dengan kebutuhan dan preferensi pasar Indonesia.
    c. Penelitian ini dapat memberikan pemahaman yang lebih baik tentang perbandingan performa yang diperoleh ketika menggunakan metode ekstraksi data TF-IDF dan *Word2Vec* pada model indoBERT
    d. Penelitian ini dapat menjadi sumber referensi yang berguna bagi peneliti yang berencana untuk melakukan penelitian terkait di masa mendatang.
2. Manfaat bagi Peneliti:
    a. Penelitian ini dapat memberikan kesempatan bagi peneliti untuk mengimplementasikan pengetahuan yang telah diperoleh selama menjalankan studi di perguruan tinggi dan mengaplikasikannya dalam konteks praktis.
    b. Penelitian ini dapat membantu peneliti dalam meningkatkan wawasan dan pengetahuan dalam memecahkan permasalahan dalam bidang *text mining*, khususnya dalam analisis sentimen dengan menggunakan model indoBERT serta perbandingan ekstraksi fitur TF-IDF dan *Word2Vec*