BAB I
PENDAHULUAN

1.1 Latar Belakang

Program *Clash of Champions* yang diselenggarakan Ruangguru merupakan salah satu program *edutainment* yang mendapat perhatian besar dari masyarakat Indonesia. Episode pertama menembus 3 juta penonton dalam tiga hari dan mencapai 10 juta tayangan dalam kurang dari tiga bulan, memicu gelombang diskusi aktif di berbagai platform media sosial [1][2]. Fenomena ini menunjukkan bahwa *Clash of Champions* tidak sekadar berfungsi sebagai media hiburan dan edukasi, tetapi juga di minati dan menarik perhatian publik secara masif dan menjadi tempat masyarakat mengekspresikan opini dan sentimen mereka [1][3]. Potensi format *edutainment* kompetisi akademik ini sebenarnya telah terbukti di tingkat global. Program serupa dari Korea Selatan, *University War*, berhasil menarik perhatian penonton internasional dan membangkitkan minat luas terhadap pendidikan STEM, namun membutuhkan beberapa musim untuk mencapai fenomena tersebut [2][4]. Hal ini menjadikan *Clash of Champions* lebih menarik untuk dikaji, program ini mencapai eksposur masif, sehingga komentar yang dihasilkan merupakan sumber data sentimen publik berbahasa Indonesia dan penting untuk dianalisis.

Sentimen publik tersebut memiliki peran penting dalam menentukan keberhasilan suatu program. Sentimen negatif yang tidak dianalisis berpotensi menurunkan dukungan audiens, sedangkan sentimen positif dapat menjadi indikator penerimaan dan kesiapan ekspansi program [5][6]. Komentar pada akun resmi Ruangguru di YouTube menjadi sumber data yang relevan karena YouTube merupakan platform video dengan jumlah pengguna terbesar [7]. Namun data yang berskala besar dan komentar yang kompleks bersifat informal, mengandung kesalahan tata bahasa, bahasa gaul, singkatan, dan slang menjadikan proses analisis secara manual tidak memungkinkan[8][9][10]. Oleh karena itu, dibutuhkan pendekatan otomatis berbasis *deep learning* yang mampu memahami kompleksitas bahasa Indonesia informal. Selain itu, tingginya volume data komentar juga menimbulkan tantangan dari sisi komputasi, di mana *machine learning* klasik cenderung tidak efisien dalam memproses data berskala besar secara cepat dan akurat[11][12]. Kompleksitas bahasa informal yang digunakan pengguna media sosial semakin memperburuk kinerja metode sederhana karena sulit dalam menangkap konteks maupun

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

makna implisit [13]. Oleh sebab itu, diperlukan pendekatan yang tidak hanya akurat, tetapi juga efisien secara komputasi serta mampu memahami karakteristik bahasa tidak terstruktur.

Beberapa penelitian terdahulu telah mengusulkan berbagai pendekatan untuk analisis sentimen, namun masing-masing masih memiliki keterbatasan. Atmaja dkk. membandingkan LSTM dan BERT pada dataset komentar *Clash of Champions* dan memperoleh akurasi tinggi (BERT 97%, LSTM 96%), namun dataset yang digunakan hanya 1.891 baris data dan belum mengeksplorasi arsitektur *hybrid* LSTM dan GRU [14]. Untuk dataset berskala lebih besar, penggunaan BERT memerlukan komputasi yang jauh lebih berat dibandingkan manfaat peningkatan akurasi yang hanya berkisar 1-2%. Khairunnisa Nasution dkk. membandingkan BiLSTM dan GRU pada ulasan produk *e-commerce* Indonesia dan menunjukkan GRU mencapai akurasi 94,07%, sedikit lebih unggul dari BiLSTM (93,61%), namun hanya menggunakan 5.400 data [15] . U. Kuriş dan Ö. C. Turna membuktikan GRU memberikan akurasi 98,67% dan LSTM 98,70% sedangkan jika digabungkan akurasi menjadi lebih baik LSTM dan GRU mencapai akurasi 99,31% Namun, penelitian tersebut hanya menggunakan dataset GPS yang relatif kecil, yaitu sebanyak 3.622 baris data [16]. Oleh karena itu, masih terdapat peluang untuk mengembangkan penelitian lebih lanjut dengan menggunakan dataset yang lebih besar dan beragam dan kompleks guna meningkatkan generalisasi model [17][10].

Meskipun LSTM mampu mengatasi permasalahan vanishing gradient dan sangat baik dalam mempelajari dependensi jangka panjang pada data sekuensial, arsitektur ini memiliki jumlah parameter yang relatif banyak karena menggunakan tiga mekanisme gerbang (forget gate, input gate, dan output gate). Semakin besar ukuran dataset dan semakin panjang urutan teks yang diproses, semakin besar pula kebutuhan komputasi, memori, dan waktu pelatihan model. [14][16][18] Untuk mengatasi keterbatasan tersebut dikembangkan Gated Recurrent Unit (GRU), yaitu penyederhanaan dari LSTM yang menggabungkan forget gate dan input gate menjadi update gate sehingga jumlah parameter menjadi lebih sedikit. Penyederhanaan tersebut membuat GRU mampu melakukan proses pelatihan lebih cepat dan lebih efisien tanpa mengurangi kemampuannya dalam mempelajari hubungan kontekstual antar kata. [15] [16] Oleh karena itu, kombinasi LSTM dan GRU menjadi alternatif yang menarik untuk memperoleh keseimbangan antara kemampuan menangkap dependensi jangka panjang dan efisiensi komputasi pada dataset berskala besar.

Berdasarkan karakteristik masing-masing arsitektur tersebut, penelitian ini mengusulkan penggunaan LSTM dan GRU untuk mengklasifikasikan sentimen komentar

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

YouTube Clash of Champions. LSTM ditempatkan pada lapisan pertama karena memiliki kemampuan yang sangat baik dalam menangkap dependensi jangka panjang melalui mekanisme cell state. [16] [19] Selanjutnya, keluaran LSTM diteruskan menuju lapisan GRU yang bertugas menyempurnakan representasi fitur secara lebih efisien melalui jumlah parameter yang lebih sedikit. [16] [19] Susunan LSTM-GRU dipilih karena penelitian Zafar dkk. menunjukkan bahwa urutan tersebut menghasilkan performa yang lebih baik dibandingkan konfigurasi sebaliknya (GRU-LSTM).[18] [20] Dengan demikian, arsitektur hybrid ini diharapkan mampu menghasilkan model yang tidak hanya akurat tetapi juga efisien dalam menangani data komentar YouTube berbahasa Indonesia.

Dataset komentar yang digunakan pada penelitian ini berjumlah 202.558 baris data (*datasets large-scale)* komentar dari akun resmi Ruangguru di YouTube jauh lebih besar data dari penelitian-penelitian sebelumnya [17] [12]. Pada dataset teks analisis sentimen yang diambil dari sosial media umumnya sering kali mengalami ketidakseimbangan kelas [21]. Untuk mengatasi permasalahan ketidakseimbangan kelas (*class imbalance*) yang dapat menimbulkan halusinasi model, penelitian ini menerapkan teknik *class weighting* pada model *Hybrid* LSTM-GRU. *Class weight* dapat meningkatkan pembelajaran pada kelas minoritas tanpa menghasilkan data sintetis. Selain menjaga keaslian data teks, metode ini juga meminimalkan risiko *noise* dan kesalahan pelabelan serta lebih efisien dalam proses pelatihan model[22][23][24].

1.2 Rumusan Masalah

Belum terdapat penelitian analisis sentimen terhadap komentar YouTube pada program *Clash of Champions* Ruangguru yang menggunakan arsitektur *hybrid* LSTM-GRU. Selain itu, dataset yang berskala besar dan tingginya volume komentar yang mencapai ratusan ribu baris data dengan karakteristik teks yang kompleks, seperti penggunaan bahasa informal, singkatan, slang, dan *imbalance* dataset yang menimbulkan tantangan dalam menghasilkan model analisis sentimen yang akurat.

1.3 Tujuan

Tujuan dari penelitian ini adalah membangun dan mengimplementasikan model *Deep Learning* dengan arsitektur *hybrid* LSTM dan GRU untuk melakukan klasifikasi sentimen (positif, negatif, dan netral) pada data komentar YouTube *Clash of Champions* Ruangguru, serta menghasilkan model yang mampu menangani data berskala besar dengan

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.

karakteristik bahasa Indonesia informal.

1.4 Manfaat

Penelitian ini memberikan kontribusi signifikan baik dari aspek akademis maupun praktis, dengan manfaat yang dapat dirasakan oleh berbagai pihak:

1. Penelitian ini diharapkan dapat menambah referensi keilmuan di bidang analisis sentimen melalui pemanfaatan data komentar media sosial untuk memahami persepsi publik terhadap konten *edutainment* terutama di bidang pendidikan.
2. Dapat digunakan oleh peneliti selanjutnya sebagai referensi untuk mengembangkan sistem yang berhubungan dengan analisis sentimen pada ulasan teks.

1.5 Batasan Masalah

Batasan masalah dari penelitian ini adalah:

1. Sumber dataset berupa data teks komentar dari sosial media YouTube pada akun resmi Ruangguru. Dataset yang dikumpulkan melalui scraping berjumlah 202.558 baris data pada season satu (95.886) dan dua (106.672) dengan rentang waktu 2024/06/28 sampai dengan 2025/09/25, dan akan dikomparasi dengan dataset dari kaggle sebanyak 85.241 baris data untuk season satu.
2. Karakteristik komentar untuk analisis dibatasi dalam berbahasa Indonesia yang bersifat informal, kesalahan tata bahasa, ambiguitas, serta mengandung singkatan dan *slang*.
3. Proses pelabelan dilakukan secara otomatis menggunakan metode *lexicon-based* dengan kamus InSet (*Indonesian Sentiment Lexicon*).
4. Representasi teks dilakukan menggunakan teknik *Integer Encoding* dengan bantuan Keras Tokenizer dan diintegrasikan dengan lapisan *Embedding Layer*.
5. Penanganan ketidakseimbangan data dilakukan dengan teknik *class weight*

© Karya Dilindungi UU Hak Cipta
1. Dilarang menyebarluaskan dokumen tanpa izin.
2. Dilarang melakukan plagiasi.
3. Pelanggaran diberlakukan sanksi sesuai peraturan UU Hak Cipta.