BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Seni merupakan bidang yang terus berkembang bersama dengan *science* dan teknologi [1]. Berbagai penemuan seperti pigmen, *printer*, kamera, komputer, *software editing*, dan *digital drawing* membuat perubahan yang signifikan terhadap bidang seni [2]. Banyak seniman yang telah merasakan dampak positif dari penemuan-penemuan tersebut dan bahkan membuat pekerjaan di bidang kreatif menjadi lebih diminati pada dua dekade belakangan ini [3].

Namun perkembangan teknologi yang pesat ini juga menimbulkan suatu kontroversi. Pada bulan Oktober tahun 2018, sebuah rumah pelelangan menjual lukisan berjudul '*Portrait of Edmond Belamy*' seharga $432,500 [4, 5, 6]. Nominal ini merupakan angka yang luar biasa, akan tetapi yang menjadi permasalahan utama di sini adalah fakta bahwa lukisan ini dibuat dengan *generative AI* yang menggunakan algoritma Generative Adversarial Networks (GANs) [5]. *Generative AI* sendiri merupakan sebuah teknologi *machine learning* yang dapat membuat suatu konten baru berupa teks, gambar, musik, atau pun video dengan menganalisis data konten lain yang telah ada sebelumnya [7].

Adanya kejadian tersebut mengakibatkan munculnya berbagai literatur yang membahas mengenai *generative AI* ini dalam berbagai perspektif [5]. Misalnya saja, penelitian yang dilakukan oleh Cetinic pada tahun 2022 [8] yang membahas mengenai permasalahan kepemilikan, hak cipta, dan etika terhadap seni yang diciptakan oleh *AI*. Kemudian Ragot melalui penelitiannya pada tahun 2020 [9] yang menjalankan eksperimen untuk mendapatkan penilaian dari partisipan terhadap lukisan yang dibuat oleh *AI* dan yang dibuat oleh tangan manusia. Ada juga penelitian terkait dengan tanggung jawab atas karya seni yang dihasilkan oleh *AI* [6]. Penelitian-penelitian ini membawa sebuah *concern* baru ke masyarakat bahwa dunia seni akan berubah secara signifikan karena adanya campur tangan *AI*.

Saat ini, terdapat lebih dari 35 kebijakan *AI* strategis nasional yang telah dikeluarkan secara global untuk mengatur permasalahan terkait *generative AI*, dan enam di antaranya merupakan perjanjian internasional. Selain itu, laporan OECD menjelaskan bahwa 50 negara telah atau sedang mengembangkan *national policy* negaranya masing-masing terkait dengan *generative AI* ini [10]. Di sisi lain, hal ini

berarti masih terdapat ratusan negara yang masih belum mengambil tindakan dan berada dalam garis abu-abu terkait dengan permasalahan ini. Pembuatan kebijakan terkait dengan *generative AI* memang masih menjadi hal yang baru yang dapat menyulitkan dan membutuhkan berbagai perspektif agar nantinya kebijakan yang dibuat dapat tepat sasaran. Maka dari itu, persepsi masyarakat akan menjadi hal yang penting bagi pembuat kebijakan [6].

Sebuah proses mengumpulkan dan menganalisis pendapat, persepsi, pemikiran, dan kesan seseorang mengenai berbagai topik, produk, subjek, dan layanan disebut sebagai analisis sentimen. Pendapat inilah yang dapat bermanfaat bagi perusahaan, pemerintah, dan individu untuk mengumpulkan informasi dan membuat suatu keputusan [11]. Ada berbagai macam algoritma yang dapat digunakan untuk melakukan analisis sentimen [12]. Namun, terdapat dua algoritma yang paling sering digunakan jika dibandingkan dengan algoritma lainnya yaitu algoritma Support Vector Machine (SVM) dan Naive Bayes (NB) [13, 14].

Penelitian yang membandingkan nilai akurasi dari kedua algoritma tersebut pernah dilakukan pada analisis sentimen komentar pengguna terkait *rating film* di IMDB dan Twitter menggunakan algoritma SVM dan NB [15]. Penelitian ini menunjukkan performa SVM yang jauh lebih baik dengan akurasi 0.84 pada *dataset* yang diambil dari Twitter dibandingkan dengan NB yang hanya memiliki akurasi sebesar 0.72 pada *dataset* yang sama. Hal yang sama juga terjadi pada analisis sentimen terkait *review* produk di Amazon menggunakan SVM dan NB. Hasilnya SVM mendapatkan nilai akurasi 84% dibandingkan dengan NB yang memiliki nilai akurasi sebesar 82.875% [16].

Penelitian untuk meningkatkan performa dari SVM juga pernah dilakukan di analisis sentimen pada *healthcare stock market* [17] dengan melakukan *hyperparameter tuning* pada parameter c, *gamma*, dan *kernel* menggunakan *grid search*. *Dataset* divektorisasi menggunakan TF-IDF sebelum digunakan pada pembuatan model SVM. Hasil dari penelitian ini menunjukkan bahwa akurasi dari SVM dapat ditingkatkan dari 81.73% menjadi 85.65%. Akurasi tersebut bisa didapatkan dengan *hyperparameter* c dengan nilai 7, *kernel* linear, dan *gamma auto*.

Penelitian-penelitian di atas memanfaatkan berbagai aplikasi seperti Twitter [15] dan Amazon [16] sebagai sumber *dataset*-nya. Selain kedua aplikasi tersebut, data yang diperlukan untuk melakukan analisis sentimen juga dapat diperoleh melalui media sosial lain seperti Youtube. Hal ini dapat dilihat pada penelitian analisis sentimen pada komentar video Youtube menggunakan berbagai algoritma [18]. Selain itu, terdapat juga penelitian yang memanfaatkan komentar di video

Youtube untuk mengetahui sentimen dan *first impression* masyarakat terhadap suatu film melalui *trailer* yang di-*upload* pada aplikasi Youtube [19]. Youtube sendiri menempati peringkat kedua sebagai *website* terpopuler di dunia dilansir melalui Alexa Internet dan Similar Web [20, 21]. Youtube memungkinkan penggunanya untuk mengetik-kan komentar pada suatu video [22]. Komentar inilah yang dapat digunakan untuk melakukan sentimen analisis [23].

Masalah yang sering terjadi terkait dengan *dataset* saat melakukan proses klasifikasi adalah distribusi *class* yang tidak merata. Data yang tidak terdistribusi secara merata ini disebut sebagai *imbalanced data*. *Imbalanced data* dapat memengaruhi proses *training* model dan dapat menyebabkan bias dalam prediksi *class*, hal ini disebabkan oleh kelas minoritas memiliki bagian yang lebih sedikit sehingga dapat dianggap sebagai *noise* atau *outlier* [24]. Salah satu solusi dari permasalahan ini pernah disinggung pada analisis sentimen untuk *review* pelanggan yaitu menggunakan Synthetic Minority Oversampling Technique (SMOTE) untuk melakukan *balancing data* [25]. Penelitian tersebut membandingkan performa SVM dibandingkan 6 algoritma lainnya untuk analisis sentimen, dan hasilnya SVM menempati peringkat pertama dengan tingkat akurasi 0.85 yang dapat ditingkatkan menjadi 0.88 dengan menggunakan SMOTE untuk *balancing data*.

Pemodelan topik merupakan pendekatan statistik untuk menemukan ”topics” yang tersembunyi dalam *corpus* pada teks [26]. Latent Dirichlet Allocation (LDA) merupakan salah satu algoritma yang biasa digunakan dalam pemodelan topik. Pemodelan topik dapat dikombinasikan dengan analisis sentimen untuk mendapatkan pemahaman lebih terkait dengan topik yang diteliti. Hal ini pernah dilakukan pada analisis sentimen terkait dengan *online education* saat pandemi Covid-19 [27]. Penelitian tersebut menghasilkan 10 kluster topik dengan 10 *keyword* yang memiliki nilai *relevance* tertinggi pada data *tweet* positif dan negatif terkait dengan *online education* di masa Covid-19.

Penelitian lain terkait pemodelan topik juga pernah dilakukan pada analisis sentimen terkait *online review* pada maskapai pesawat [28]. Hasil dari penelitian tersebut menampilkan berbagai *keyword* yang terkait dengan 6 kluster topik utama yaitu ’*In-flight meal*’, ’*Entertainment*’, ’*Seat class*’, ’*Seat comfort*’, ’*Singapore airline*’, dan ’*Staff Service*’ dari maskapai pesawat Singapore Airlines. Sebagai lanjutan dari penelitian-penelitian yang telah disebutkan di atas, penelitian ini menggunakan algoritma SVM untuk melakukan analisis sentimen serta mengekstrak kluster topik dari hasil prediksi sentimen menggunakan pemodelan topik dengan algoritma LDA sehingga diperoleh gambaran tentang topik pada

sentimen positif, negatif, dan netral yang dipikirkan oleh pengguna Youtube terkait dengan *generative AI*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan diatas, adapun rumusan masalah dari penelitian ini yaitu sebagai berikut.

1. Bagaimana cara mengimplementasikan algoritma SVM dan LDA untuk melakukan analisis sentimen dan pemodelan topik terkait dengan *generative AI* pada platform Youtube?
2. Berapa nilai *accuracy*, *precision*, *recall*, dan *f1-score* dari algoritma Support Vector Machine dalam melakukan analisis sentimen pada komentar video Youtube yang terkait dengan *generative AI*?

1.3 Batasan Permasalahan

Adapun beberapa batasan masalah yang ada pada penelitian ini yaitu sebagai berikut.

1. *Dataset* diambil dari komentar Youtube di video terkait dengan *generative AI* dengan total sebanyak 21.877 komentar pengguna.
2. *Dataset* terakhir kali dikumpulkan pada tanggal 25 Februari 2024.

1.4 Tujuan Penelitian

Penelitian ini dilakukan dengan tujuan sebagai berikut.

1. Mengimplementasikan algoritma SVM dan LDA untuk melakukan analisis sentimen dan pemodelan topik terkait dengan *generative AI* pada platform Youtube.
2. Menghitung nilai *accuracy*, *precision*, *recall*, dan *f1-score* dari algoritma Support Vector Machine dalam melakukan analisis sentimen pada komentar video Youtube yang terkait dengan *generative AI*.

1.5 Manfaat Penelitian

Manfaat yang didapatkan dari hasil penelitian ini adalah menjadi salah satu bahan pertimbangan bagi pemerintah untuk mengambil keputusan terkait dengan pembuatan perundang-undangan yang berkaitan dengan *generative AI* berdasarkan pendapat positif dan negatif dari masyarakat di Youtube.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini berisikan latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan terkait dengan judul penelitian.

- Bab 2 LANDASAN TEORI
  Bab ini berisikan landasan teori terkait dengan judul penelitian. Landasan teori yang dijelaskan pada bab ini adalah *generative AI*, analisis sentimen, textblob, algoritma Support Vector Machine (SVM), algoritma SMOTE, *confusion matrix*, pemodelan topik, algoritma Latent Dirichlet Allocation (LDA), PyLDAvis, *text pre-processing*, TF-IDF, serta *data cleaning* yang terdiri dari *tokenization*, *removing stop word*, *lemmatization*, dan *labeling*

- Bab 3 METODOLOGI PENELITIAN
  Bab ini berisikan metodologi penelitian yang digunakan pada penelitian ini. Metodologi penelitian ini terdiri dari alur penelitian, studi literatur, pengumpulan data, *text pre-processing*, *data cleaning*, *tokenization*, *removing stop word*, *lemmatization*, *labeling*, TF-IDF, *data splitting*, analisis sentimen dengan SVM, pengujian model SVM, dan pemodelan topik dengan LDA dan PyLDAvis serta diakhiri dengan dokumentasi.

- Bab 4 HASIL DAN DISKUSI
  Bab ini berisikan hasil dan pembahasan dari penelitian yang telah dilakukan sesuai dengan metode penelitian pada bab 3.

- Bab 5 KESIMPULAN DAN SARAN
  Bab ini berisikan kesimpulan yang menjawab tujuan dari penelitian dan saran yang dapat diberikan untuk penelitian selanjutnya.