

BAB I

PENDAHULUAN

A. Latar Belakang

Kemajuan teknologi informasi telah mendorong meningkatnya penggunaan media sosial sebagai sarana komunikasi, informasi, dan diskusi publik. Salah satu platform yang banyak digunakan adalah *YouTube*, di mana pengguna dapat mengunggah video, memberikan komentar, serta berdiskusi tentang berbagai isu yang sedang terjadi. Salah satu topik yang menarik perhatian publik akhir-akhir ini adalah peluncuran DANANTARA di Indonesia. Peluncuran ini menimbulkan berbagai reaksi dari masyarakat, baik yang mendukung maupun yang mengkritik, yang disampaikan melalui berbagai media, termasuk komentar di *YouTube*.

Reaksi yang muncul dari masyarakat terhadap peluncuran DANANTARA disebabkan oleh berbagai faktor, seperti persepsi terhadap manfaat program ini, kekhawatiran terhadap dampaknya, serta tingkat pemahaman masyarakat mengenai tujuan dan implementasinya. Sebagian masyarakat melihat peluncuran ini sebagai langkah positif yang dapat membawa perubahan atau kemajuan tertentu, sementara yang lain merasa tidak percaya atau bahkan khawatir terhadap konsekuensi yang mungkin ditimbulkan, baik dari segi sosial, ekonomi, maupun politik. Pembentukan Badan Pengelola Investasi Daya Anagata Nusantara (BPI Danantara) memang membawa dampak positif dalam hal percepatan pertumbuhan ekonomi melalui pengelolaan aset negara yang lebih terintegrasi dan profesional, serta mampu menarik investasi ke sektor-sektor strategis seperti infrastruktur dan energi. Namun di sisi lain, kehadiran badan ini juga menimbulkan kekhawatiran terkait potensi penyalahgunaan wewenang dan intervensi politik yang dapat mengganggu objektivitas serta efektivitas pengelolaan investasi [1]. Ketidakseimbangan informasi, perbedaan latar belakang, serta opini yang berkembang di media turut memperbesar ruang diskusi publik, terutama di platform seperti *YouTube* yang memungkinkan penyampaian pendapat secara langsung dan terbuka. Hal ini menjadikan komentar-komentar di *YouTube* sebagai cerminan dari beragam sudut pandang masyarakat yang perlu dianalisis lebih dalam untuk memahami respons publik secara komprehensif.

Analisis sentimen adalah metode yang efektif dalam mengidentifikasi dan memahami sikap, opini, dan emosi yang terkait dengan suatu topik tertentu di kalangan masyarakat. Teknik ini memanfaatkan pengolahan data teks untuk menentukan apakah sentimen yang terkandung dalam teks tersebut bersifat positif, netral, atau negatif. Hal ini membantu dalam mengekstrak informasi tentang evaluasi atau opini pribadi yang diberikan oleh pengguna internet melalui media sosial atau platform lainnya [2]. Namun, dalam praktiknya, kualitas hasil analisis sentimen sangat dipengaruhi oleh data yang tidak seimbang (*imbalanced data*), di mana jumlah komentar dengan sentimen tertentu jauh lebih banyak dibandingkan sentimen lainnya. Ketidakeimbangan data ini sering terjadi dalam *Dataset* komentar publik dan dapat menyebabkan menurunnya akurasi model klasifikasi, termasuk algoritma *Naïve Bayes*.

Naïve Bayes merupakan salah satu algoritma pembelajaran mesin yang populer dalam analisis sentimen karena sifatnya yang sederhana, cepat, dan efisien dalam memproses data teks. *Naïve Bayes* telah terbukti menjadi salah satu pendekatan yang efektif dalam analisis sentimen. Metode ini berfokus pada pemodelan probabilitas kemunculan kata-kata atau frasa dalam konteks sentimen tertentu (positif, negatif, atau netral). Dengan menggunakan data latih yang berisi teks-teks yang sudah diberi label sentimen, model *Naïve Bayes* dapat mempelajari pola dan korelasi antara kata-kata dengan sentimen tertentu. Kemudian, model ini dapat digunakan untuk mengklasifikasikan teks baru ke dalam kategori sentimen yang sesuai [3]. Algoritma ini didasarkan pada Teorema Bayes dengan asumsi independensi antar fitur, yang meskipun sederhana, telah terbukti efektif dalam berbagai tugas klasifikasi teks, termasuk analisis sentimen. Selain itu, *Naïve Bayes* mampu menangani *Dataset* besar dengan waktu komputasi yang relatif cepat dibandingkan dengan algoritma lain seperti *Support Vector Machine* (SVM) atau *deep learning*. Oleh karena itu, *Naïve Bayes* dipilih sebagai metode utama dalam penelitian ini untuk menganalisis sentimen komentar terkait peluncuran DANANTARA di Indonesia.

Permasalahan yang berpengaruh dalam proses klasifikasi adalah adanya data *imbalance*. Data *imbalance* merupakan kondisi ketika banyaknya label kelas data latih yang satu lebih sedikit daripada banyaknya label kelas data yang lainnya.

Ketika jumlah data dalam satu kelas jauh lebih banyak dibandingkan kelas lainnya, model cenderung bias terhadap kelas mayoritas dan kurang akurat dalam mengklasifikasikan kelas minoritas [4].

Salah satu solusi untuk mengatasi *imbalanced data* adalah dengan menggunakan metode *undersampling*, seperti *Random Undersampling* dan *Tomek Link*. *Random Undersampling* (RUS) merupakan metode *sampling* secara acak memilih sampel di kelas mayoritas sehingga sejajar dengan sampel dari kelas minoritas [5]. Sementara itu, *Tomek Link* adalah metode yang mengidentifikasi pasangan data dari kelas berbeda yang memiliki jarak *Euclidean* paling dekat dalam ruang fitur. Pasangan ini dianggap berada di area batas keputusan yang ambigu, sehingga dapat dihapus (baik dari kelas mayoritas maupun minoritas) sebagai langkah pembersihan data setelah pemrosesan awal [6].

Penelitian [7] menganalisis sentimen pengguna terhadap aplikasi Kalender Bali dengan mengelompokkan ulasan menjadi kategori positif dan negatif. Algoritma *K-Nearest Neighbour* (KNN) digunakan untuk melakukan klasifikasi sentimen guna mengukur tingkat kepuasan pengguna secara umum. Untuk meningkatkan kinerja klasifikasi, teknik *Tomek Link* diterapkan bersama dengan KNN. Data dianalisis berdasarkan dua periode, yaitu sebelum dan sesudah pandemi COVID-19, sebagai respons terhadap lonjakan ulasan negatif yang muncul setelah pembaruan aplikasi selama masa pandemi. Pada periode sebelum COVID-19, algoritma KNN menunjukkan akurasi sebesar 94,3% tanpa penggunaan *Tomek Link* dan 93,7% saat metode ini diterapkan, masing-masing dengan nilai $K=3$ dan $K=5$. Sementara itu, pada periode setelah COVID-19, akurasi yang diperoleh meningkat dari 86,0% menjadi 87,2% dengan penerapan *Tomek Link* dan parameter $K=9$. Peningkatan akurasi sebesar 1,2% pada data pasca-COVID menunjukkan efektivitas *Tomek Link* dalam menangani distribusi data yang tidak seimbang. Hasil ini menegaskan bahwa meskipun KNN memiliki performa yang cukup baik, penerapan *Tomek Link* mampu meningkatkan akurasi klasifikasi secara signifikan, khususnya ketika data cenderung tidak seimbang.

Penelitian [8] menganalisis sentimen pengguna Twitter terhadap Undang-Undang Cipta Kerja melalui hashtag #UUCIPTAKERJA. Sentimen diklasifikasikan menjadi positif dan negatif menggunakan algoritma *Naïve Bayes*. Hasil awal

menunjukkan adanya ketidakseimbangan antara nilai *Recall* dan spesifisitas akibat distribusi data latih yang tidak seimbang. Untuk mengatasi hal ini, digunakan metode *Random Undersampling* yang secara acak mengurangi jumlah data pada kelas mayoritas. Metode ini dipilih karena mampu meningkatkan performa klasifikasi dan mempercepat waktu eksekusi.

Penelitian [9] mengkaji peningkatan kinerja metode *Random Forest* pada klasifikasi data diabetes yang tidak seimbang dengan menerapkan kombinasi *Smote-Tomek Link*. Data yang digunakan berasal dari Kaggle dengan 768 data dan delapan atribut *input*. Hasil validasi silang *10-Fold* menunjukkan bahwa *Random Forest* dengan *Smote-Tomek Link* memberikan hasil lebih baik dibandingkan dengan hanya menggunakan *Smote*. Penerapan *Tomek Link* dalam proses ini berperan dalam membersihkan data hasil *oversampling*, sehingga menghasilkan akurasi 86,4%, sensitivitas 88,2%, presisi 82,3%, dan skor F1 sebesar 85,1%. Hal ini menunjukkan bahwa integrasi *Tomek Link* dapat meningkatkan kualitas klasifikasi pada data yang tidak seimbang.

Penerapan *Random Undersampling* dan *Tomek Link* dalam penelitian ini bertujuan untuk mengatasi ketidakseimbangan data dalam analisis sentimen komentar *YouTube* terkait peluncuran DANANTARA di Indonesia. Dengan menerapkan teknik ini, diharapkan model *Naïve Bayes* dapat bekerja lebih optimal dalam mengklasifikasikan sentimen positif, negatif, dan netral. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan metode analisis sentimen yang lebih akurat dan andal untuk memahami opini publik terhadap isu-isu ekonomi dan teknologi di Indonesia.

B. Perumusan Masalah

1. Apa urgensi penelitian analisis sentimen terhadap komentar *YouTube* pada isu kebijakan publik seperti peluncuran DANANTARA, serta apa keistimewaan penelitian ini yang berfokus pada penanganan ketidakseimbangan data?
2. Mengapa data komentar *YouTube* cenderung memiliki distribusi sentimen yang tidak seimbang, khususnya pada isu kebijakan publik, dan bagaimana kondisi tersebut memengaruhi proses analisis sentimen?

3. Bagaimana hasil analisis sentimen diimplementasikan ke dalam sebuah aplikasi sebagai alat bantu untuk mengamati dan mengevaluasi hasil penanganan data tidak seimbang?
4. Bagaimana pengaruh kombinasi metode *Random Undersampling* dan *Tomek Link* terhadap performa algoritma *Naïve Bayes* dalam analisis sentimen komentar *YouTube* terkait peluncuran DANANTARA?

C. Batasan Penelitian

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Penelitian ini menggunakan data komentar *YouTube* yang relevan dengan topik peluncuran DANANTARA, dengan total 12.232 komentar yang dikumpulkan dari dua video pada dua *channel YouTube*, yaitu video berjudul "**SURAT TERBUKA UNTUK DANANTARA**" yang diunggah oleh *channel* Raymond Chin, dan video berjudul "**Memahami Danantara dengan Mudah**" dari *channel Malaka Project*. Data diambil dan dikumpulkan pada tanggal 12 Agustus 2025, dalam format CSV.
2. Komentar yang dianalisis hanya dibatasi pada komentar berbahasa Indonesia dan bahasa Inggris. Komentar berbahasa asing selain bahasa Inggris akan dihapus atau difilter dalam tahap *Preprocessing*. Jika terdapat komentar berbahasa Inggris, maka akan diterjemahkan terlebih dahulu ke dalam bahasa Indonesia dalam tahap *Preprocessing* agar sesuai dengan pendekatan leksikon yang digunakan.
3. Penelitian ini membahas dua teknik *undersampling*, yaitu *Random Undersampling* dan *Tomek Link*, sebagai upaya penanganan terhadap ketidakseimbangan data. Teknik *balancing* data lainnya tidak dibahas dalam ruang lingkup penelitian ini.
4. Analisis sentimen dibatasi pada tiga kategori, yaitu sentimen positif, negatif, dan netral.
5. Pelabelan data sentimen dilakukan menggunakan pendekatan berbasis leksikon dengan memanfaatkan InSet (*Indonesia Sentiment Lexicon*) yang mencakup kosakata bernuansa positif dan negatif dalam bahasa Indonesia.

6. Menggunakan Tesaurus Bahasa Indonesia sebagai kamus atau daftar kata-kata Bahasa Indonesia yang memiliki makna, sinonim dan antonim yang digunakan dalam proses *Preprocessing* dalam penanganan negasi.
7. Ekstraksi fitur menggunakan metode TF-IDF (*Term Frequency-Inverse Document Frequency*) dengan jumlah fitur maksimal sebanyak 10.000.
8. Kinerja model akan dievaluasi menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dengan menggunakan validasi silang *K-Fold* di mana nilai *K* ditetapkan sebesar 10.
9. Proses prapemrosesan (*Preprocessing*) data teks dalam penelitian ini tidak mencakup penanganan emotikon, sehingga emotikon yang terdapat dalam komentar tidak digunakan sebagai fitur dalam analisis sentimen.
10. Penelitian ini difokuskan pada kondisi data yang memiliki ketidakseimbangan label sentimen (*imbalanced data*), sehingga penerapan dan evaluasi metode penanganan ketidakseimbangan data dilakukan apabila distribusi label sentimen tidak seimbang.
11. *Website* aplikasi yang dikembangkan dalam penelitian ini bersifat *proof-of-concept* dan dijalankan secara lokal pada *environment* pengembangan (*localhost*), tanpa dipublikasikan pada layanan *hosting* komersial seperti *shared hosting*, *VPS*, atau *cloud platform*. Keputusan ini diambil karena fokus utama penelitian terletak pada pengembangan dan evaluasi model *machine learning* menggunakan algoritma *Naive Bayes* dengan metode *Random Undersampling* dan *Tomek Link* untuk menangani data tidak seimbang, bukan pada aspek *deployment* dan operasional sistem berbasis *web* secara *production*. Selain itu, terdapat keterbatasan *resources* untuk biaya *hosting* dan *maintenance* aplikasi dalam jangka panjang yang berada di luar lingkup penelitian skripsi. Aplikasi ini telah memenuhi fungsinya sebagai alat bantu demonstratif untuk memvisualisasikan kelayakan teknis dan alur kerja sistem analisis sentimen, mulai dari tahap *crawling data YouTube*, *preprocessing* teks, pelabelan dengan *Inset Lexicon*, pembobotan TF-IDF, hingga klasifikasi dan evaluasi performa model. Potensi implementasi aplikasi dalam lingkungan *production* untuk

penggunaan aktual dibahas lebih lanjut dalam bagian saran penelitian sebagai rekomendasi pengembangan masa depan.

D. Tujuan Penelitian

1. Menjelaskan urgensi analisis sentimen terhadap komentar *YouTube* pada isu kebijakan publik seperti peluncuran DANANTARA, serta mengidentifikasi keistimewaan penelitian yang berfokus pada penanganan ketidakseimbangan data.
2. Menganalisis karakteristik distribusi sentimen pada data komentar *YouTube* terhadap isu kebijakan publik serta menjelaskan penyebab terjadinya ketidakseimbangan data dan dampaknya terhadap proses analisis sentimen.
3. Mengimplementasikan hasil analisis sentimen ke dalam sebuah aplikasi sebagai alat bantu untuk mengamati dan mengevaluasi hasil penanganan data tidak seimbang.
4. Penelitian ini bertujuan untuk mengetahui pengaruh metode *Tomek Link-Random Undersampling* terhadap performa algoritma *Naïve Bayes* dalam analisis sentimen komentar *YouTube* terkait peluncuran DANANTARA.

E. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik dari segi akademis maupun praktis, di antaranya:

a. Manfaat Akademis:

1. Menambah wawasan di bidang *machine learning*, khususnya dalam penerapan algoritma *Naïve Bayes* dengan pendekatan *undersampling* menggunakan kombinasi *Tomek Link* dan *Random Undersampling*.
2. Memberikan kontribusi terhadap pengembangan teknik penanganan data tidak seimbang (*imbalanced data*) untuk meningkatkan performa algoritma klasifikasi dalam analisis sentimen.
3. Menjadi referensi bagi penelitian selanjutnya yang ingin mengkaji efektivitas metode *undersampling* dalam meningkatkan akurasi dan kestabilan model klasifikasi pada data sentimen.

b. Manfaat Praktis:

1. Membantu memahami persepsi masyarakat terhadap peluncuran DANANTARA melalui analisis komentar pengguna di platform *YouTube*.
2. Memberikan wawasan bagi pemerintah atau pihak terkait mengenai tanggapan publik terhadap peluncuran program atau kebijakan, sehingga dapat dijadikan bahan evaluasi dalam pengambilan keputusan.

