BAB 1

PENDAHULUAN

1.1 Latar Belakang

Seiring kemajuan dunia digital, informasi kini semakin mudah diakses melalui portal dan media sosial. Media sosial berfungsi sebagai wadah bagi masyarakat untuk menyampaikan opini, baik berupa pujian maupun kritik, termasuk ujaran kebencian dan hoaks yang kerap memicu perdebatan. Kini, media sosial telah menjadi bagian penting dalam interaksi di dunia maya serta dalam upaya mencari informasi. Di dalamnya, konten biasanya disajikan dalam bentuk teks yang dikelompokkan berdasarkan isi dan topik tertentu. Dengan popularitas platform media sosial, memantau sentimen publik secara real-time menjadi sangat krusial bagi bisnis dan organisasi pemerintah di masa kini (Xu*et al.*, 2022).

Menurut laporan berjudul "Digital Indonesia 2024" oleh We Are Social dan Meltwater, dua organisasi yang berfokus pada riset digital dan analisis media sosial, tercatat ada 611,3 juta pengguna X (sebelumnya dikenal sebagai Twitter) di seluruh dunia pada April 2024. Indonesia menempati urutan keempat sebagai negara dengan jumlah pengguna X terbanyak. Indonesia memiliki 24,85 juta pengguna, menjadikannya salah satu pasar media sosial terbesar di dunia. Angka ini menunjukkan popularitas X di tanah air dan

menunjukkan peran pentingnya dalam percakapan sosial dan tren digital. Sebagai salah satu media sosial yang banyak digunakan di Indonesia, X adalah *platform* yang dinamis yang memungkinkan pengguna untuk berbagi perspektif mereka secara *real-time* mengenai berbagai isu yang tengah populer. Melalui fitur *tweet*, *retweet*, suka, dan balasan, pengguna dapat mengekspresikan perasaan, opini, dan tanggapan mereka terhadap topik yang sering kali rumit dan kontroversial. Hal ini menjadikan X sebagai sumber data yang berpotensi besar untuk dianalisis dan diklasifikasi lebih lanjut (Basar *et al.*, 2022), terutama dalam mengidentifikasi sentimen publik secara lebih mendalam.

Analisis sentimen merupakan salah satu bentuk aplikasi *text mining*. Tujuan utamanya adalah untuk mengevaluasi dan memahami opini yang terkandung di dalam dokumen teks (Wahyuningsih *et al.*, 2024). Proses ini pada dasarnya merupakan masalah klasifikasi teks, dengan sentimen sebagai kategori utamanya. Shagun *et al*. (2024) menyatakan bahwa sebagai komponen dari *Natural Language Processing* (*NLP*), analisis sentimen memungkinkan komputer untuk mengenali emosi manusia baik positif, negatif, maupun netral dari sejumlah besar data tekstual. Wawasan yang diperoleh dari analisis sentimen memiliki berbagai aplikasi di berbagai sektor, memberikan informasi berharga untuk pengambilan keputusan yang lebih terinformasi dan berdampak. Analisis sentimen terhadap isu di media sosial belakangan ini telah menjadi alat yang krusial untuk memahami opini publik dalam berbagai bidang, termasuk pendidikan (Alawi dan Bozkurt, 2024).

Pendekatan *supervised machine learning* telah terbukti efektif dalam analisis sentimen, dengan berbagai algoritma seperti *Support Vector Machine* (*SVM*), *Random Forest*, *Maximum Entropy*, *Neural Networks*, Naïve Bayes, dan *K-Nearest Neighbor* (*KNN*) yang sering digunakan. Pemilihan algoritma yang paling cocok bergantung pada kompleksitas data dan kebutuhan interpretasi. Jika akurasi menjadi prioritas utama, maka *Random Forest*

adalah pilihan terbaik, seperti yang diungkapkan oleh Wahyuningsih *et al.* (2024). *Random Forest* merupakan teknik *ensemble learning* yang didasarkan pada penggabungan beberapa algoritma *Decision Tree* untuk meningkatkan akurasi prediksi dan stabilitas model (Bilal *et al.*, 2016). Hal ini sejalan dengan penelitian terdahulu oleh Ghulam *et al.* (2021) yang membandingkan kinerja dari tiga algoritma *machine learning*, yaitu *Logistic Regression*, Naïve Bayes, dan *Random Forest* untuk mengklasifikasikan argumen siswa. Hasilnya menunjukkan bahwa *Random Forest* lebih unggul dalam menangani data yang lebih kompleks karena algoritma ini mampu mengelola fitur-fitur yang saling berkaitan dengan baik. Lebih lanjut, penelitian oleh Fitri (2020) membandingkan algoritma Naïve Bayes, *Random Forest*, dan *Support Vector Machine* (*SVM*) dalam analisis sentimen terhadap aplikasi Ruangguru. Dari ketiga algoritma tersebut, *Random Forest* memberikan performa terbaik dengan tingkat akurasi tertinggi mencapai 97,16%, mengungguli algoritma lainnya. Penelitian lain oleh Neogi *et al.* (2021) yang membahas sentimen Twitter terkait protes petani di India, juga mendukung temuan ini. Dalam penelitian tersebut, *Random Forest* terbukti menghasilkan akurasi terbaik yaitu 95,51%, mengungguli metode klasifikasi lainnya seperti Naïve Bayes, *Decision Tree*, dan *Support Vector Machine* (*SVM*).

Menurut beberapa penelitian, salah satu masalah utama dalam proses klasifikasi sentimen adalah kualitas teks yang tidak konsisten pada media sosial. Seperti yang diungkapkan oleh Agarwal *et al.* (2014), penulisan opini di media sosial sering kali menghadapi kendala karena adanya tweet yang sulit dibaca. Hal ini disebabkan oleh beberapa faktor, seperti penulisan kata yang disingkat, penggunaan bahasa *gaul* atau *slang*, kesalahan ketik, serta gaya penulisan yang tidak baku. Hambatan ini menunjukkan pentingnya representasi ulang kata sebelum mengambil keputusan mengenai emosi seseorang. Setelah tahap ini selesai, kata-kata tersebut akan dikategorikan untuk menentukan apakah termasuk dalam perasaan positif, netral, atau negatif. Penelitian sebelumnya oleh Antinasari *et al.* (2017) menunjukkan bahwa proses normalisasi bahasa menggunakan algoritma *Levenshtein*

Distance mampu meningkatkan akurasi secara signifikan dalam klasifikasi opini film pada dokumen Twitter berbahasa Indonesia menggunakan metode Naïve Bayes. Namun, hasil berbeda ditemukan pada penelitian Manuaba *et al.* (2022), di mana penerapan algoritma *Levenshtein Distance* pada metode *Support Vector Machine* (*SVM*) untuk analisis sentimen data provider layanan internet pada Twitter tidak memberikan peningkatan akurasi yang signifikan. Hal ini menunjukkan bahwa efektivitas algoritma *Levenshtein Distance* dalam proses normalisasi bahasa dapat bervariasi tergantung pada model klasifikasi yang digunakan.

Penelitian ini bertujuan mengeksplorasi sekaligus menguji sejauh mana penerapan algoritma *Levenshtein Distance* dapat meningkatkan performa *Random Forest* dalam klasifikasi sentimen berbahasa Indonesia terkait kebijakan terbaru program beasiswa Lembaga Pengelola Dana Pendidikan (LPDP). Sebelumnya, penerima beasiswa LPDP diwajibkan untuk kembali ke Indonesia setelah menyelesaikan studi di luar negeri. Namun, rencana Kemendikti Saintek untuk mengkaji ulang program ini memunculkan spekulasi mengenai perubahan kebijakan sekaligus meningkatkan sorotan terhadap isu kebermanfaatan program LPDP. Tak lama kemudian, perhatian publik semakin terfokus setelah Menteri Pendidikan Tinggi, Sains, dan Teknologi (Mendikti Saintek) pada tahun 2024, Satryo Soemantri Brodjonegoro, memberikan kelonggaran kepada beberapa penerima beasiswa untuk tidak lagi diwajibkan pulang ke tanah air setelah menyelesaikan studi mereka, sebagaimana disampaikan dalam artikel Kompas.com pada 5 November 2024. Perubahan ini memicu pro dan kontra, baik di kalangan masyarakat maupun pejabat negara. Polemik ini memicu perdebatan karena dianggap tidak berkontribusi pada kemajuan negeri. Kasus ini menjadi *viral* di media sosial X, dengan tagar 'LPDP' yang sempat menjadi *trending*. Hal tersebut memperkuat alasan pemilihan X sebagai objek dalam penelitian ini. *Platform* ini tidak hanya digunakan oleh calon dan penerima beasiswa LPDP untuk menyampaikan kritik, dukungan, maupun aspirasi mereka, tetapi juga menjadi ruang bagi masyarakat luas

untuk turut serta memberikan opini. Keterlibatan publik yang masif dalam diskusi kebijakan ini menjadikan X sebagai representasi sentimen masyarakat yang paling dinamis dan *real-time*. Karakteristik X yang terbuka, cepat menyebarkan informasi, serta berbasis teks yang mudah dianalisis, menjadikannya sumber data utama yang sangat potensial dalam mengkaji persepsi publik terhadap isu-isu aktual seperti perubahan kebijakan LPDP.

Oleh karena itu, penelitian ini bertujuan untuk menganalisis opini masyarakat di media sosial X terkait kebijakan terbaru LPDP, dengan memanfaatkan algoritma *Levenshtein Distance* untuk proses normalisasi teks, serta *Random Forest* untuk klasifikasi sentimen. Dengan pendekatan ini, diharapkan dapat memberikan gambaran yang lebih jelas mengenai sentimen masyarakat terhadap program beasiswa LPDP dan dampak kebijakan baru tersebut terhadap persepsi publik. Dengan memanfaatkan keunggulan *Random Forest* yang mampu menangani data dengan dimensi tinggi dan memiliki mekanisme bawaan untuk mencegah *overfitting*, diharapkan pendekatan ini mampu menghasilkan akurasi yang lebih baik dibandingkan metode sebelumnya.

1.2 Perumusan Masalah

Berdasarkan latar belakang tersebut maka peneliti melakukan pengembangan dengan rumusan masalah pada penelitian yaitu:

1. Bagaimana performa algoritma *Levenshtein Distance* dalam tahap normalisasi teks terhadap akurasi klasifikasi sentimen terkait isu kebijakan baru Program Beasiswa LPDP menggunakan model *Random Forest*?

2. Bagaimana sentimen pengguna media sosial X terhadap isu kebijakan baru Program Beasiswa LPDP menggunakan *Random Forest*?

3. Bagaimana hasil analisis sentimen terhadap isu kebijakan baru Program

Beasiswa LPDP dapat membantu pemangku kepentingan memahami kebutuhan dan harapan masyarakat?

1.3 Tujuan Penelitian

1. Menganalisis pengaruh penerapan algoritma *Levenshtein Distance* pada tahap normalisasi teks terhadap hasil klasifikasi sentimen menggunakan model *Random Forest*.
2. Mengetahui sentimen positif, netral, atau negatif pengguna media sosial X terhadap Program Beasiswa LPDP.
3. Memberikan wawasan yang dapat membantu pemangku kepentingan dalam memahami dan memenuhi kebutuhan serta harapan masyarakat terkait kebijakan baru Program Beasiswa LPDP.

1.4 Batasan Masalah

Terdapat beberapa batasan masalah dalam penelitian yaitu sebagai berikut:

1. Sumber data yang digunakan berasal dari hasil *scraping* data X dengan *tweet* yang diunggah pada rentang tanggal 31 Oktober 2024 – 28 Februari 2025, menggunakan kata kunci 'LPDP'.
2. Data *tweet* yang digunakan merupakan teks dalam bahasa Indonesia.
3. Klasifikasi sentimen dibagi menjadi tiga, yaitu sentimen positif, netral, dan negatif.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat bermanfaat bagi pihak yang terkait, yaitu:

1. Bagi Penulis

   Penelitian ini memberikan kesempatan untuk memperluas wawasan dan keterampilan dalam analisis sentimen serta penerapan teknik *data mining* pada data berbahasa Indonesia, sehingga meningkatkan kompetensi di bidang ini.

2. Bagi Program Studi Matematika

   Penelitian ini dapat menjadi tolak ukur keberhasilan mahasiswa dalam menerapkan konsep matematika terapan, khususnya di bidang analisis data dan *machine learning*, serta berkontribusi pada pengembangan keilmuan di lingkungan akademik.

3. Bagi Pembaca atau Umum

   Penelitian ini menyediakan informasi yang relevan mengenai sentimen masyarakat terhadap kebijakan baru Program Beasiswa LPDP dan dapat menjadi referensi bagi pemangku kepentingan dalam merumuskan kebijakan yang responsif, serta memberikan landasan bagi penelitian lebih lanjut di bidang yang sama atau bidang terkait.