

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Opini sangat penting dalam hampir semua aktivitas manusia karena mereka merupakan faktor kunci yang memengaruhi perilaku kita. Setiap kali kita perlu membuat keputusan, kita ingin mengetahui pendapat orang lain. Dalam dunia nyata, bisnis dan organisasi selalu ingin mengetahui pendapat konsumen atau publik tentang produk dan layanan mereka. Konsumen individu juga ingin mengetahui pendapat pengguna lain mengenai suatu produk sebelum membelinya. Di masa lalu, ketika seseorang membutuhkan opini, dia bertanya kepada teman dan keluarga. Ketika organisasi atau bisnis membutuhkan opini publik atau konsumen, mereka melakukan survei, jajak pendapat, dan diskusi kelompok. Memperoleh opini publik dan konsumen telah lama menjadi bisnis besar itu sendiri bagi perusahaan pemasaran [1]. Namun pada zaman modern ini, memperoleh informasi dari konsumen dapat dilakukan dengan efisiensi dan efektivitas yang lebih tinggi. Sebab berkembangnya ilmu analisis sentimen, merupakan sebuah teknik yang digunakan untuk mengidentifikasi dan mengklasifikasikan opini atau perasaan yang terkandung dalam teks. Dengan menganalisis sejumlah besar data teks, seperti ulasan produk, posting media sosial, dan komentar online, perusahaan dapat memperoleh wawasan berharga tentang persepsi pelanggan terhadap produk atau layanan mereka [2].

Maka dari itu, sentimen analisis berperan penting dalam mendapatkan informasi sentimen dari sebuah teks konsumen, dalam kasus ini penulis menggunakan data ulasan restoran cepat saji McDonald's di Amerika Serikat. Sentimen analisis diterapkan untuk menganalisa emosi dan perspektif penulis teks ulasan tersebut. Sentimen analisis atau penggalian opini merupakan bentuk pengaplikasian dari pemrosesan bahasa, linguistik, dan analisis teks agar informasi subjektif dari teks ulasan tersebut dapat ditemukan dan diekstrak. Bidang teknologi ini menganalisa sekaligus memprediksi informasi tersembunyi dari sebuah teks. Informasi tersembunyi ini memberikan masukan yang bermanfaat mengenai intensi, gaya, dan kemungkinan konsumen tersebut. Sentimen analisis dikhususkan mengkategorisasi teks dalam tingkat subjektif dan objektif [3]. Sebelum penulis memilih algoritma-algoritma yang akan dibandingkan performanya dalam menjalankan analisis sentimen, ada penelitian-penelitian sejenis dengan penggunaan algoritma lain sebagai bahan referensi penulis dalam meneliti.

Sebagai contoh, Penelitian ini mengembangkan sebuah metode untuk mengidentifikasi sentimen (positif, negatif, atau netral) dalam ulasan pelanggan restoran Pakistan dan mengelompokkan ulasan tersebut ke dalam kategori tertentu. Data yang digunakan berasal dari postingan pengguna di grup media sosial "SWOT" (komunitas hobi kuliner yang terkenal di Pakistan). Model yang dikembangkan menggunakan teknik pemrosesan bahasa alami (NLP) untuk mengolah teks ulasan. Dengan menggunakan empat algoritma yang mengklasifikasikan kategori (rasa makanan, kesepadanan nilai dengan harga, suasana, dan pelayanan) memperoleh akurasi, yaitu Regresi Logistik (86%), *Naïve*

Bayes (84%), *Support Vector Machine* (89%), dan *Random Forest* (96%), diuji dalam penelitian ini. Hasil terbaik diperoleh dengan menggunakan algoritma *Random Forest*, yang mencapai akurasi di atas 90% dalam analisis sentimen dan klasifikasi kategori. Meskipun *Random Forest* mendapatkan hasil terbaik, perlu dicatat bahwa tiga algoritma lainnya mencapai akurasi di atas 90% dalam melakukan analisis sentimen dengan tiga rasio training dan testing yang berbeda yaitu, Regresi Logistik (92%), *Naïve Bayes* (92%), *Support Vector Machine* (93%), dan *Random Forest* (95%) persentase sentimen analisis yang ditampilkan di sini menggunakan rasio training dan testing 75:25 (pada dua rasio training dan testing lain juga menghasilkan nilai akurasi di atas 90%) sama seperti rasio pada klasifikasi kategori, sedangkan empat persentase pertama pada algoritma di atas adalah persentase klasifikasi kategori [4]. Dengan hasil penelitian tersebut, ada penelitian pembandingan sebagai tambahan referensi penulis.

Selanjutnya, Dalam penelitian ini, menunjukkan dataset baru dalam domain restoran untuk bahasa Uzbekistan, dengan 8210 ulasan, diberi label positif atau negatif, yang dikumpulkan dari Google Maps menggunakan URL semua lokasi di ibu kota Uzbekistan yaitu Tashkent, dan diberi label sesuai dengan skor bintangnya. Kemudian, menerapkan langkah-langkah pra-pemrosesan penuh pada dataset yang berkontribusi pada peningkatan akurasi model. Analisis lebih lanjut dari dataset yang dikumpulkan ditunjukkan dengan evaluasi menggunakan teknik machine learning dan *deep learning*. Hasil akurasi terbaik (91%) pada dataset diperoleh menggunakan model Regresi Logistik dengan *n-grams* kata dan karakter. Model ini lebih baik dibandingkan model lain yang digunakan beserta parameternya adalah

Regresi Logistik dengan *n-grams* kata, Regresi logistik dengan *n-grams* karakter, *Support Vector Machine* dengan Kernel Linear, *Recurrent Neural Network* dengan dan tanpa Word Embeddings, dan Convolutional Neural Networks (Multichannel) [5]. Berdasarkan hasil penelitian tersebut, ada penelitian pembandingan tambahan sebagai referensi bagi penulis.

Seperti, Penelitian ini menemukan bahwa kebanyakan pelanggan restoran di Labuan Bajo memiliki pandangan positif terhadap makanan, produk, layanan, harga, suasana, dan lokasi restoran. Namun, pengelola restoran perlu memperhatikan preferensi pelanggan yang berbeda-beda, terutama berdasarkan jenis perjalanan mereka. Selain itu, penelitian ini juga menunjukkan bahwa metode pemrosesan data ulasan yang menggabungkan teknik SMOTE Upsampling dengan algoritma k-NN memberikan hasil yang paling akurat, mencapai 99,27%. Metode ini lebih baik daripada metode lain seperti *Naïve Bayes*, *Decision Tree*, dan *Support Vector Machine* [6]. Dengan penelitian-penelitian yang membandingkan model-model algoritma tersebut dalam menganalisis sentimen pada ulasan-ulasan di sektor makanan dan minuman, penulis mendapatkan wawasan dan memilih model-model algoritma yang akan digunakan.

Dengan demikian, setelah penulis melakukan pertimbangan, pada proses analisis sentimen ini penulis menggunakan 3 model algoritma yaitu *Naïve Bayes*, SVM, dan *Random Forest* yang nantinya berdasarkan hasil evaluasi tersebut akan dibandingkan performanya. Klasifikasi *Naïve Bayes* sendiri adalah sebuah keluarga klasifikasi yang cukup mirip dengan model linier. Namun, mereka cenderung lebih cepat dalam proses pelatihan. Namun, dengan efisiensi ini model *Naïve Bayes*

sering kali memberikan kinerja generalisasi yang sedikit lebih buruk dibandingkan dengan classifier linier seperti LogisticRegression dan LinearSVC. Alasan mengapa model *Naïve Bayes* sangat efisien adalah karena mereka mempelajari parameter dengan melihat setiap fitur secara individual dan mengumpulkan statistik sederhana per kelas dari setiap fitur. Sedangkan *Random Forest* pada dasarnya adalah kumpulan pohon keputusan, di mana setiap pohon sedikit berbeda dari yang lainnya. Gagasan di balik *random forest* adalah bahwa setiap pohon mungkin melakukan pekerjaan yang cukup baik dalam memprediksi, tetapi kemungkinan akan *overfitting* pada sebagian data. Jika kita membangun banyak pohon, yang semuanya bekerja dengan baik namun *overfitting* dengan cara yang berbeda, kita dapat mengurangi tingkat *overfitting* dengan merata-ratakan hasil mereka. Pengurangan *overfitting* ini, sambil tetap mempertahankan kekuatan prediktif dari pohon-pohon tersebut, dapat ditunjukkan dengan matematika yang ketat. Untuk menerapkan strategi ini, kita perlu membangun banyak pohon keputusan. Setiap pohon harus melakukan pekerjaan yang memadai dalam memprediksi target, dan juga harus berbeda dari pohon-pohon lainnya [7].

Selanjutnya, algoritma terakhir yaitu *Support Vector Machine* (SVM) yang membawa revolusi baru dalam komunitas machine learning. SVM didasarkan pada kerangka teori yang kuat dan terbukti menjadi pengklasifikasi yang efisien. SVM menawarkan pendekatan yang teliti terhadap masalah pembelajaran mesin. SVM membangun solusinya berdasarkan subset dari input pelatihan. SVM telah banyak digunakan untuk tugas klasifikasi, regresi, pengenalan pola, dan pengurangan fitur. Gagasan utama SVM adalah mengurangi kesalahan generalisasi dengan

memaksimalkan margin dari sebuah hyperplane. SVM didasarkan pada landasan teori pembelajaran statistik yang kuat. SVM telah berhasil diterapkan dalam berbagai bidang, seperti pengenalan digit tulisan tangan, pengenalan objek, identifikasi pembicara, pengenalan pola, kategorisasi teks, biomedis, peramalan deret waktu keuangan, dan banyak lagi [8]. Itulah beberapa penjelasan mengenai model-model algoritma yang akan digunakan pada penelitian ini. Untuk ulasannya sendiri diambil dari restoran cepat saji Amerika Serikat.

Yakni McDonald's adalah jaringan restoran cepat saji terbesar di dunia, melayani lebih dari 69 juta pelanggan setiap hari di lebih dari 100 negara di lebih dari 40.000 gerai di tahun 2021 [9]. Penulis memilih McDonald's karena memiliki volume ulasan yang sangat besar di berbagai platform media sosial. Analisis sentimen terhadap ulasan McDonald's dapat memberikan wawasan berharga bagi perusahaan dan juga menjadi kasus studi yang menarik untuk penelitian akademik. Lalu berdasarkan ulasan konsumen pada dataset yang penulis gunakan (dataset ini terakhir kali diperbarui pada tahun 2023 diunggah oleh Nidula Elgiryewithana pada Kaggle), ada beberapa masalah pada pelayanan restoran cepat saji McDonald's di Amerika Serikat yaitu kualitas makanan yang jelek, servis yang lambat, kesalahan pesanan, kebersihan yang tidak memadai, karyawan yang tidak ramah, dan inkonsistensi kualitas pengalaman seperti yang dilaporkan oleh American Customer Satisfaction Index organization (ACSI) [10].

Dengan mempertimbangkan berbagai faktor, analisis perbandingan kinerja algoritma *Naïve Bayes*, SVM, dan *Random Forest* dalam mendeteksi sentimen pada ulasan McDonald's menunjukkan bahwa pemilihan algoritma yang tepat sangat

krusial untuk mencapai akurasi dan relevansi yang optimal. Hasil penelitian ini dapat menjadi acuan bagi pengembangan sistem deteksi sentimen yang lebih baik di masa depan, baik dalam konteks bisnis maupun penelitian.

Berdasarkan latar belakang yang ada di atas, maka penulis memutuskan untuk mengambil judul **“ANALISIS PERBANDINGAN KINERJA ALGORITMA NAÏVE BAYES, SVM, DAN RANDOM FOREST DALAM MENDETEKSI SENTIMEN PADA ULASAN MCDONALD’S”**

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang di atas maka perumusan masalah yang dapat dilakukan adalah “Bagaimana kinerja algoritma *Naïve Bayes*, *SVM*, dan *Random Forest* dalam mengklasifikasikan sentimen konsumen pada ulasan McDonald's?”

1.3 BATASAN MASALAH

Adapun batasan masalah yang penulis tetapkan untuk menghindari terjadinya perluasan pembahasan dan di luar ruang lingkup masalah pada penelitian ini, adalah sebagai berikut:

1. Data ulasan diambil secara khusus dari kaggle yang dataset pada platform Google Reviews.
2. Ulasan yang dianalisis hanya dalam bahasa Inggris.
3. Ulasan hanya diambil dari cabang McDonald's di wilayah Amerika Serikat.

4. Perbandingan dilakukan pada tiga algoritma, yaitu *Naïve Bayes*, *Support Vector Machine (SVM)*, dan *Random Forest*.
5. Data teks akan melalui proses *preprocessing* meliputi tokenisasi, *case folding*, dan *stop word removal* menggunakan NLTK.
6. Sentimen dikategorikan menjadi tiga kelas: positif, negatif, dan netral.
7. Kinerja algoritma akan dievaluasi menggunakan metrik akurasi, precision, recall, F1-score, dan AUC (*Area Under Curve*).

1.4 TUJUAN DAN MANFAAT PENELITIAN

1.4.1 Tujuan Penelitian

Tujuan penelitian ini, adalah sebagai berikut:

1. Membandingkan kinerja: Membandingkan akurasi, presisi, recall, dan F1-score dari algoritma *Naïve Bayes*, *SVM*, dan *Random Forest* dalam mengklasifikasikan sentimen (positif, negatif, netral) pada ulasan McDonald's.
2. Menentukan algoritma terbaik: Mengidentifikasi algoritma mana yang paling efektif dalam mendeteksi sentimen pada dataset ulasan McDonald's.
3. Mengetahui keterbatasan: Mengidentifikasi keterbatasan dari masing-masing algoritma dalam konteks analisis sentimen pada ulasan McDonald's.

1.4.2 Manfaat Penelitian

Manfaat penelitian ini, adalah sebagai berikut:

1. Pengembangan sistem rekomendasi: Hasil penelitian dapat digunakan untuk mengembangkan sistem rekomendasi yang lebih akurat bagi konsumen McDonald's, berdasarkan sentimen mereka terhadap produk dan layanan.
2. Pengambilan keputusan bisnis: Perusahaan dapat memanfaatkan hasil penelitian untuk membuat keputusan bisnis yang lebih baik, seperti meningkatkan kualitas produk atau layanan yang mendapatkan umpan balik negatif.
3. Pengembangan model analisis sentimen: Hasil penelitian dapat berkontribusi pada pengembangan model analisis sentimen yang lebih baik dan efisien untuk berbagai domain.

1.5 SISTEMATIKA PENELITIAN

Untuk memberikan gambaran umum mengenai keseluruhan proyek penelitian dapat dilihat melalui sistematika penelitian yang meliputi:

BAB 1 : PENDAHULUAN

Pada bab ini merupakan bab pendahuluan yang mengemukakan latar belakang masalah, perumusan masalah, batasan masalah, tujuan, dan manfaat penelitian, dan sistematika penulisan.

BAB II : LANDASAN TEORI

Pada bab ini, berisikan tentang kajian-kajian dan konsep umum mengenai, NLP (*Natural Language Processing*), analisis sentimen, algoritma klasifikasi *Naïve Bayes*, *SVM*, *Random Forest*, *preprocessing* text, evaluasi kinerja (metrik evaluasi), dan studi kasus sebagai pendukung proses penelitian yang bersumber dari berbagai media seperti, buku, jurnal, artikel, internet, maupun sumber-sumber terkait lainnya.

BAB III : METODOLOGI PENELITIAN

Pada bab ini akan menguraikan tentang tahapan proses yang dilakukan selama mengerjakan penelitian, metode atau pendekatan yang digunakan, serta alat bantu atau *tools* yang digunakan dalam perancangan ini, baik dalam hal *hardware* maupun *software*.

BAB IV : ANALISIS

Pada bab ini peneliti akan mempresentasikan hasil-hasil yang diperoleh model analisis sentimen dengan penerapan algoritma *Naïve Bayes*, *SVM*, dan *Random Forest* pada data ulasan McDonald's. Hasil-hasil tersebut akan disajikan dalam bentuk tabel, grafik, atau visualisasi lainnya yang mudah dipahami. Selain itu, peneliti juga akan melakukan interpretasi terhadap hasil-hasil

tersebut. Analisis mendalam akan dilakukan untuk membandingkan kinerja dari ketiga algoritma, mengidentifikasi algoritma mana yang memberikan hasil terbaik dalam mendeteksi sentimen pada ulasan McDonald's, dan menganalisis faktor-faktor yang mempengaruhi kinerja masing-masing algoritma. Dengan kata lain, bab ini bertujuan untuk menjawab pertanyaan penelitian yang diajukan di awal dan memberikan bukti empiris untuk mendukung kesimpulan yang diambil.

BAB V : PENUTUP

Pada bab ini berisikan kesimpulan dari hasil analisis sentimen model, serta saran-saran yang disampaikan terkait dengan hasil penelitian ini.