

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi membuat penggunaan media sosial semakin meningkat sebagai sarana komunikasi dan berbagi informasi. Platform berbasis teks seperti X menghasilkan data cuitan dalam jumlah besar yang dapat dimanfaatkan untuk menganalisis karakteristik pengguna. Data tersebut digunakan dalam berbagai bidang, seperti analisis opini publik, rekomendasi konten, serta pemasaran digital yang membutuhkan pemahaman karakter pengguna [1].

Salah satu informasi penting dalam analisis media sosial adalah jenis kelamin pengguna. Namun, data gender tidak selalu tersedia karena banyak pengguna tidak mencantumkan identitas lengkap atau menggunakan akun anonim. Kondisi ini menyulitkan proses identifikasi karakteristik pengguna jika hanya mengandalkan profil akun. Oleh karena itu, diperlukan pendekatan lain yang mampu memperkirakan gender berdasarkan pola bahasa pada teks cuitan [2].

Beberapa penelitian menunjukkan adanya perbedaan kecenderungan penggunaan bahasa antara laki-laki dan perempuan dalam komunikasi digital [3][4]. Cuitan pengguna laki-laki umumnya lebih berkaitan dengan aktivitas, topik tertentu, atau informasi langsung, sedangkan pengguna perempuan lebih sering menggunakan bahasa yang ekspresif dan personal. Perbedaan pola tersebut menunjukkan bahwa karakteristik linguistik dapat dimanfaatkan sebagai dasar dalam klasifikasi gender berbasis teks.

Seiring dengan temuan bahwa karakteristik linguistik dapat dimanfaatkan untuk klasifikasi gender berbasis teks, perkembangan NLP memungkinkan analisis teks dilakukan secara otomatis menggunakan machine learning. Penelitian ini menggunakan Logistic Regression, Random Forest, dan Support Vector Machine karena umum digunakan dalam klasifikasi teks dengan pendekatan berbeda [5]. Oleh karena itu, penelitian ini bertujuan menganalisis pola bahasa pengguna X serta membandingkan kinerja ketiga algoritma dalam klasifikasi gender berbasis cuitan.

1.2 Rumusan Masalah

Informasi jenis kelamin pengguna pada media sosial sering kali tidak tersedia atau tidak dicantumkan secara jelas, sehingga menyulitkan proses identifikasi karakteristik pengguna berdasarkan data demografis. Ketergantungan pada data profil menjadi tidak efektif ketika pengguna menggunakan akun anonim atau tidak melengkapi identitasnya. Di sisi lain, meskipun terdapat perbedaan kecenderungan penggunaan bahasa antara laki-laki dan perempuan dalam komunikasi digital, pemanfaatan karakteristik linguistik pada teks cuitan sebagai dasar identifikasi gender masih belum sepenuhnya terstruktur dan teruji secara konsisten.

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini ditetapkan untuk memperjelas ruang lingkup penelitian agar pembahasan tetap terfokus dan tidak meluas. Adapun batasan masalah pada penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya menggunakan data berupa teks cuitan (tweet) pengguna media sosial X (sebelumnya Twitter) tanpa mempertimbangkan metadata lain seperti jumlah pengikut, lokasi pengguna, waktu unggahan, atau interaksi (like, retweet, dan reply).
2. Klasifikasi jenis kelamin pengguna dibatasi hanya pada dua kelas, yaitu laki-laki (male) dan perempuan (female). Data dengan label lain seperti brand, unknown, atau data tanpa label tidak disertakan dalam penelitian.
3. Dataset yang digunakan merupakan dataset publik yang diperoleh dari Kaggle, dengan jumlah data akhir sebanyak 1.026 cuitan yang telah melalui proses seleksi dan pelabelan ulang secara manual.
4. Metode ekstraksi fitur teks yang digunakan dalam penelitian ini hanya Term Frequency-Inverse Document Frequency (TF-IDF) dengan kombinasi unigram dan bigram, tanpa menggunakan metode ekstraksi fitur lain seperti Word2Vec, GloVe, atau pendekatan berbasis deep learning.
5. Algoritma klasifikasi yang digunakan dibatasi pada tiga metode, yaitu Logistic Regression, Random Forest, dan Support Vector Machine, tanpa melakukan pengujian pada algoritma klasifikasi lainnya.

6. Evaluasi kinerja model dilakukan menggunakan metrik akurasi, precision, recall, dan F1-score, dengan skema K-Fold Cross Validation, tanpa melakukan analisis lebih lanjut.
7. Penelitian ini tidak membahas aspek etika, identitas asli pengguna, maupun implikasi sosial dari klasifikasi gender, dan hanya berfokus pada performa model klasifikasi berdasarkan konten teks cuitan

1.4 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Menganalisis perbedaan pola penggunaan bahasa antara pengguna laki-laki dan perempuan berdasarkan kata dominan dan kata khas pada teks cuitan.
2. Membandingkan kinerja algoritma Logistic Regression, Random Forest, dan Support Vector Machine dalam melakukan klasifikasi gender pengguna berdasarkan isi teks cuitan.

1.5 Manfaat Penelitian

1.5.1 Manfaat Teoritis

Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan kajian text mining dan machine learning, khususnya pada bidang klasifikasi gender berbasis teks cuitan media sosial. Hasil penelitian ini dapat menjadi referensi tambahan terkait pemanfaatan fitur linguistik dan metode ekstraksi fitur TF-IDF dalam membedakan karakteristik bahasa berdasarkan jenis kelamin pengguna. Selain itu, penelitian ini juga dapat digunakan sebagai bahan perbandingan dan acuan bagi penelitian selanjutnya yang membahas topik klasifikasi teks atau analisis gender pada media sosial.

1.5.2 Manfaat Praktis

1. Penelitian ini dapat dimanfaatkan sebagai referensi teknis dalam pengembangan sistem klasifikasi gender berbasis teks pada platform media sosial atau aplikasi analisis data berbasis bahasa alami.
2. Hasil penelitian ini dapat membantu pihak yang berkepentingan, seperti peneliti, analis data, atau pengembang sistem, dalam memilih algoritma

klasifikasi yang sesuai untuk data teks berdimensi tinggi, khususnya pada kasus klasifikasi gender.

3. Informasi mengenai perbedaan pola bahasa antara pengguna laki-laki dan perempuan dapat dimanfaatkan sebagai dasar analisis perilaku pengguna, misalnya dalam pengelompokan konten atau analisis kecenderungan komunikasi di media sosial.
4. Selain itu, penelitian ini juga berpotensi dimanfaatkan dalam berbagai bidang analisis media sosial, seperti analisis opini publik, sistem rekomendasi konten, serta pemasaran digital, khususnya untuk memahami karakteristik pengguna ketika informasi profil tidak tersedia.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran secara umum mengenai isi dan alur pembahasan pada setiap bab yang disajikan. Adapun sistematika penulisan dalam skripsi ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan skripsi. Bab ini bertujuan untuk memberikan gambaran awal mengenai permasalahan dan arah penelitian yang dilakukan.

BAB II TINJAUAN PUSTAKA

Bab ini memuat kajian pustaka dan penelitian terdahulu yang relevan dengan topik penelitian. Selain itu, pada bab ini juga dibahas dasar-dasar teori yang mendukung penelitian, seperti konsep data mining, text mining, pra-pemrosesan teks, ekstraksi fitur TF-IDF, algoritma klasifikasi Logistic Regression, Random Forest, dan Support Vector Machine, evaluasi model, serta k-fold.

BAB III METODE PENELITIAN

Bab ini menjelaskan metode penelitian yang digunakan, meliputi objek penelitian, sumber dan karakteristik data, tahapan pengolahan data, pra-pemrosesan teks, ekstraksi fitur, metode klasifikasi yang digunakan, serta teknik evaluasi kinerja

model.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil pengolahan dan analisis data yang telah dilakukan. Pembahasan meliputi hasil eksplorasi data awal (EDA), analisis pola bahasa berdasarkan gender, hasil ekstraksi fitur TF-IDF, serta hasil evaluasi dan perbandingan kinerja algoritma klasifikasi yang digunakan.

BAB V PENUTUP

Bab ini berisi kesimpulan yang diperoleh dari seluruh rangkaian penelitian serta saran yang dapat dijadikan sebagai bahan pertimbangan untuk penelitian selanjutnya.

