# BAB II
# TINJAUAN PUSTAKA

## 2.1. Landasan Teori

### 2.1.1. Data terpilah

Menurut (KPPA, 2015) Data terpilah Ialah data yang diklasifikasi menurut aspek atau variabel agama, pendidikan, tempat, ras, atau jenis kelamin. Data terpilah menurut jenis kelamin digunakan sebagai masukan mendasar dalam mengerjakan analisis gender ialah salah satu tindakan perlu yang harus dilakukan dalam mengerjakan Pengarus Utamaan Gender (PUG). Dengan kian berkembangnya pemahaman isu gender dalam pembangunan, semakin dirasakan juga perlunya ketersediaan akan kualitas data yang dapat memberikan gambaran isu gender.

### 2.1.2. Identifikasi jenis kelamin berdasarkan nama

Menurut (Muzdalifah Muhammadun, 2016) identifikasi jenis kelamin dapat diungkapkan secara leksikal, yaitu dalam bentuk satuan lingual berupa indikasi laki-laki dan perempuan untuk manusia, serta indikasi jantan dan betina untuk binatang. idnetifikasi jenis kelamin secara fonemis (bunyi yang membedakan arti) berbentuk sufiks-a untuk gambaran maskulin, seperti putra, dewa dan sufiks-i untuk gambaran feminin, seperti putri dan dewi. Dalam bahasa indonesia juga menandai kedua gambaran ini secara morfemis berbentuk penambahan sufiks –man, -wan, -in sebagai indeks maskulin dan sufiks –wati dan –at sebagai indeks feminin. Contoh sufiks di atas bersumber dari bahasa asing, yaitu bahasa arab dan bahasa sansekerta.

### 2.1.3. Data Preprocessing

Preprocessing data menurut (Hambardzumyan, 2021) merupakan salah satu tugas penambangan data yang paling banyak yang meliputi penyiapan dan transformasi data menjadi bentuk yang sesuai dengan prosedur penambangan. Tujuan preprocessing data untuk mengurangi ukuran data, menemukan hubungan antara data, normalisasi data, hapus outlier, dan ekstrak fitur untuk data. Ini

mencakup beberapa teknik seperti data cleaning, integrasi, transformasi, dan reduksi. berikut keterangan dari masing-masing tahapan (Hambardzumyan, 2021):

1) Data Cleaning merupakan langkah pertama teknik preprocessing data yang digunakan untuk menemukan nilai yang hilang, data noise yang halus, mengenali outlier dan benar tidak konsisten. Data kotor ini akan berpengaruh pada miming prosedur dan menyebabkan keluaran yang tidak dapat diandalkan dan buruk.
2) Data Integration merupakan teknik yang bekerja dengan menggabungkan data dari multi dan berbagai sumber daya menjadi penyimpanan data yang konsisten.
3) Data Transformation merupakan teknik pengubah data menjadi bentuk yang cocok untuk diproses, sepertihalnya menyesuaikan data nilai kedalam rentang tertentu seperti antara 0-1, ini berguna untuk teknik seperti klasifikasi.

### 2.1.4. Fitur Ekstraksi

Teknik ekstraksi fitur yang umum ialah *Bag-of-Word* (BoW), *Term Frequency* (TF), *Word2Vec* dan *Term Frequency-Inverse Document Frequency* (TF-IDF). Fitur ekstraksi umumnya mengubah isi teks menjadi daftar terpisah atau token standar (Kowsari et al., 2019). Secara teknis kata-kata dipetakan dalam kosakata berbasis indeks, untuk menyederhanakan representasi internal token. Namun, token atau indeks harus tetap direpresentasikan dalam bentuk yang dapat dicerna oleh mesin, yaitu bentuk vektor (Kowsari et al., 2019).

#### 1) Bag-of-word

Model bag-of-words adalah representasi dokumen tekstual dan disederhanakan dari bagian yang dipilih, berdasarkan kriteria tertentu, seperti frekuensi kata. Teknik BoW digunakan di beberapa area seperti NLP, computer vision, Bayesian filtering, serta klasifikasi dokumen dan informasi machine learning (Kowsari et al., 2019).

Bow penadahan teks semacam kalimat maupun dokumen, dihitung bagaikan sekumpulan kata. Dalam sistem BoW. Kata-kata ini dibuat ke dalam sebuah matriks dan bukan merupakan kalimat yang terstruktur dalam segi tata bahasa maupun hubungan integritas serta pengabaian antara kata-kata. Diversitas dihitung digunakan ke dalam fokus pada dokumen. Hal ini dapat dicapai dengan membuat

*Term Document Matrix* (TDM). Ini hanyalah sebuah matriks dengan istilah sebagai baris dan nama dokumen sebagai kolom, hitungan frekuensi kata sebagai sel matriks. Sklearn menyediakan fungsi yang baik di bawah feature_ekstraksi.teks untuk mengonversi kumpulan dokumen teks menjadi matriks jumlah kata (Swamynathan, 2019). Pada tabel x merupakan contoh *Term Document Matrix* dengan data dokumen berikut:

V1: Text mining is to find useful information from text

V2: Useful information is mined from the text

V3: Dark came

**Tabel 2.1** Tahap vektorisasi bag-of-word

| | text | mining | is | to | find | useful | information | from | text | mined | dark | came |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| V2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| V3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

Pada tabel 2.1 merupakan tahap *vector space model* pada *bag-of-word* yang mana setiap dokumen dari korpus direpresentasikan sebagai vektor multidimensi, setiap istilah mewakili satu dimensi ruang vektor. Dalam kasus bobot biner mengambil nilai 0 dan 1 dalam mewakili keberadaan kata, 1 mencerminkan adanya kata dan 0 sebaliknya, pengambilan setiap kata ditandakan dengan 1-gram atau setiap adanya *blank space* maka akan mewakili satu dimensi vektor. Tahap selanjutnya dihitung nilai frekuensi dari setiap kata dalam setiap dokumen, semakin banyak frekuensikata maka semakin penting bagi dokumen.

**2) N-gram**

Dalam merepresentasikan BoW, salah satu masalah yang dihadapi ada pada hilangnya suatu konteks dan juga selalu berfokuskan pada setiap kata yang diperlihatkan secara terpisah, dalam hal ini tidak menghiraukan keterkaitan kata dalam suatu kalimat, pada kata sesudah dan sebelumnya yang membuat hilangnya informasi semantik atau bermakna pada saat kalimat dipisah menjadi sekumpulan kata mandiri (Khomsah & Agus Sasmito Aribowo, 2020).

*N-Gram* merumuskan gabungan kata-kata tetangga atau panjang huruf n ke dalam teks tertentu. Sebentuk *N-Gram* mewakili kumpulan kata atau karakter n (ditunjukkan sebagai gram menunjukkan tata bahasa) yang mengikuti satu sama lain. N-Gram dimanfaatkan dalam memprediksi kata selanjutnya beralaskan N-1 kata sebelumnya (Mulyani et al., 2021).

Kata yang tercantum dimanfaatkan dalam menciptakan indeks, terhadap seberapa sering kata-kata yang menyertai satu sama lain. N-gram dapat diindikasikan ke dalam rumus berikut (Mulyani et al., 2021):

$$Ngramsk = X - (N - 1) \quad (2.1)$$

### 2.1.5. Supervised learning

Menurut (Retnoningsih & Pramudita, 2020) Supervised learning merupakan suatu program dalam pembuatan AI. Diucap supervised sebab dalam pendekatan ini, machine learning dilatih dalam menandai pola antara input data dan label output.

Menurut (Fansyuri, 2020) Supervised learning merupakan sebuah algoritma program pada machine learning yang bermaksud dalam mengumpulkan suatu data ke data yang sudah ada melalui cara melatih data yang sudah ada tersebut dan terdapat variabel yang ditargetkan. Beberapa metode supervised learning yang sering digunakan dalam kasus klasifikasi antara lain yaitu Logistik Regresi, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), dan Decission Trees.

### 2.1.6. Logistic Regression

Algoritma *logistic regression* (LR) merupakan teknik pembelajaran statistik yang digunakan untuk tugas klasifikasi. Meskipun nama pengklasifikasi ini memiliki kata regresi didalamnya, itu digunakan untuk menghasilkan output biner (Kowsari et al., 2019). Metode ini dimanfaatkan pada saat variabel predictor (y) memiliki skala kategorik yang terdiri dari dua (biner) atau lebih. Oleh sebab itu LR dibuat untuk meyakinkan bahwa apapun perkiraan yang diperoleh, akan berada di antara 1 dan 0 (Reviantika et al., 2021).

$$P(Y) = \frac{e(a+bx1+cX2+\cdots+nXn)}{1+e(a+bx1+cX2+\cdots+nXn)} \quad (2.2)$$

### 2.1.7. Multinomial Naive Bayes

Algoritma Multinomial Naive Bayes (MNB) merupakan salah satu variasi dari naive bayes. MNB, mengingat konteks kelas dan mengabaikan semua dependensi antar atribut, serta memperkirakan semua atribut saling bersangkutan satu sama lain. Banyaknya dokumen (n) masuk ke dalam (k) kategori dimana k ∈ {c1, c2, ..., ck} sebagai kelas prediksi, dan keluarannya berupa c ∈ C (Kowsari et al., 2019). Rumus ada pada persamaan ().

$$P(c \mid d) = \frac{P(c) \prod w \in d \; P(d \mid c) nwd}{P(d)} \quad (2.3)$$

$^{nwd}$ dilambangkan dengan berapa kali kata w muncul dalam dokumen, dan P(w|c) adalah probabilitas dari mengamati kata w diberikan kelas c. Rumus pada persamaan ().

$$P(w \mid c) = \frac{1 + \sum d \in Dc \; nw}{k + \sum w' \sum d \in Dc \; nw' \; d} \quad (2.4)$$

### 2.1.8. Evaluasi

Evaluasi dapat digunakan menggunakan suatu ukuran tertentu, yang mana salah satunya yaitu confusion matrix banyak digunakan dalam machine learning sebagai penentuan model klasifikasi supervised learning (Hasnain et al., 2020). Struktur persegi dari sebuah confusion matrix diwakili melalui baris dan kolom,

dimana baris adalah kelas aktual dari instance, dan kolom adalah kelas yang diprediksi. Untuk klasifikasi biner, Confusion matrix direpresentasikan sebagai matriks 2 x 2. Empat representasi dari hasil proses klasifikasi confusion matrix yaitu 'true positive' (TP), 'true negative' (TN), 'false positive' (FP), dan 'false neagtive' (FN) (Hasnain et al., 2020).

**Tabel 2.2** Tabel yang dihasilkan confusion matrix

| | | **Prediksi** | |
|---|---|---|---|
| | | **TRUE** | **FALSE** |
| **Aktual** | **TRUE** | TP | FN |
| | **FALSE** | FP | TN |

Dari tabel *confusion matrix* di atas dapat dihitung accuracy, precision, recall, dan f1-score dengan rumus:

**Tabel 2.3** Tabel perhitungan confusion matrix

| **Matrik** | **Keterangan** | **Rumus** | **Persamaan** |
|---|---|---|---|
| *Accuracy* | Persentase yang diprediksi benar. | $\frac{(TP + TN)}{(TP + FP + TN + FN)}$ | (2.5) |
| *Recall* | Persentase perbandingan antara TP dengan banyaknya data yang diprediksi salah tetapi mempunyai nilai aktual benar (FN). | $\frac{TP}{(TP + FN)}$ | (2.6) |
| *Precision* | Persentase perbandingan antara TP dengan banyaknya data yang diprediksi benar tetapi mempunyai nilai aktual salah (FP). | $\frac{TP}{(TP + FP)}$ | (2.7) |
| *F1-Score* | Persentase rata-rata harmonik dari nilai recall dan precision. | $2 * \frac{(presisi * recall)}{(presisi + recall)}$ | (2.8) |

Pada tabel 2.3 merupakan rumus nilai evaluasi dalam pengevaluasian model, terdapat persamaan *accuracy*, *recall*, *precision*, dan *F1-score*.

## 2.2. Studi Literatur

Penelitian tentang klasifikasi jenis kelamin berdasarkan nama dengan metode fitur ekstraksi dan algoritma *supervised learning* bukan pertama kalinya dilakukan, terdapat beberapa penelitian sebelumnya yang telah menyelesaikan studi kasus dengan berbagai macam metode yang bergantung pada keperluan penelitian.

(Septiandri, 2017) dengan judul Predicting the Gender of Indonesian Names mengklasifikasikan jenis kelamin pada nama Indonesia menggunakan Character-Level Long-Short Term Memory dibandingkan dengan Naive Bayes, Logistic Regression, dan XGBoost menggunakan fitur n-grams. Hasil penelitian menunjukkan bahwa kinerja Naive Bayes dan Regresi Logistik yang terbaik diperoleh dari 3-gram, dan kinerja terbaik dari XGBoost adalah dari 2-gram. Saat menggunakan teknik character-level long-short term memory, mereka mampu mengklasifikasikan jenis kelamin lebih akurat daripada kemampuan Regresi Logistik. Persentase akurasi naik dari 85,28% menjadi 92,25% dalam kasus nama lengkap, sedangkan penggunaan nama depan hanya menghasilkan tingkat akurasi 90,65%.

Penelitian lainnya (Yuenyong & Sinthupinyo, 2020) dengan judul Gender Classification of Thai Facebook Usernames, Metode atau Fitur yang diusulkan, termasuk tokenisasi kata, klasifikasi bagian ucapan, frekuensi karakter, dan karakter substring dapat mencapai hasil yang baik. Hasil eksperimen menunjukkan bahwa penggunaan tokenisasi kata untuk semua nama pengguna mencapai tingkat akurasi dasar 65,81%, tetapi model gabungan mencapai peningkatan kinerja dengan tingkat akurasi 91,75%.

Penelitian lainnya (Ho Huong et al., 2022) dengan judul A Computational Linguistic Approach for Gender Prediction Based on Vietnamese Names, penelitian dilakukan pada nama-nama orang vietnam berjumlah 3 juta data, terdapat 4 tahapan kondisi penelitian yaitu penggunaan 1-3 gram dengan nama lengkap, penggunaan 1-3 gram tanpa nama keluarga, menggunakan nama tengah, dan penggunaan

kombinasi 1 gram dengan TF. Didapatkan hasil akurasi tertinggi ada pada algoritma LR dengan kondisi 1 gram dan tanpa nama keluarga, mendapatkan 90.9% akurasi.

Penelitian lainnya (Rego et al., 2021) dengan judul Predicting Gender of Brazilian Names Using Deep Learning, Memeriksa dan menerapkan model jaringan saraf dalam umpan maju dan berulang, seperti MLP, RNN, GRU, CNN, dan BiLSTM, untuk mengklasifikasikan jenis kelamin melalui nama depan. Dataset nama Brasil digunakan untuk melatih dan mengevaluasi model. Menganalisis matriks akurasi, recall, presisi, dan confusion matrix untuk mengukur kinerja model. Hasil menunjukkan bahwa prediksi jenis kelamin dapat dilakukan dari strategi ekstraksi fitur melihat nama sebagai satu set string. Beberapa model secara akurat memprediksi jenis kelamin lebih dari 90% kasus. Model berulang mengatasi model feedforward dalam masalah klasifikasi biner ini.

Penelitianlainnya (Khan et al., 2020) dilakukan pada data kostumer ecommerce yang berjumlah 15.000, dilakukan 4 tahap kondisi penelitian yaitu pendekatan pembobotan yang berbeda serta pendekatan augmentasi sesi, termasuk dekomposisi ID unik, pembuatan jendela konteks, dan hierarki identik dengan n-fold cross-validation. Dan didapatkan hasil akhir F1-score tertinggi ada pada pengujian identical hierarchy dengan cross validasi 3 yaitu mendapatkan 0.9754 untuk perempuan dan 0.9102 untuk laki-laki, sedangkan untuk persentase rata-rata f1-score didapatkan 0.9427 macro dan 0.9613 micro.

(Zhao & Kamareddine, 2018) menyajikan metode untuk menganalisis data online untuk perbedaan gender dalam bidang ilmu komputer di Inggris, Malaysia, dan Cina. Pada penelitian menyempurnakan alat prediksi gender untuk nama depan yang membantu melengkapi data online dengan lebih akurat dalam karakter dua bahasa yang berbeda. Sistem dapat menampilkan hasilnya kepada pengguna secara langsung pada grafik dinamis. Metode ini berguna bagi peneliti sosial untuk mengolah data besar saat membuat prediksi jenis kelamin nama depan. Kami melakukan eksperimen dengan alat kami dalam menganalisis perbedaan gender dalam ilmu komputer di Inggris, Malaysia, dan China.

(Zakia, 2020) Hasil pengujian menunjukkan bahwa nilai ketetanggaan yang optimal ialah k=3. Pada k=3 nilai akurasi, precision, recall dan F-Measure dari rerata 10-Fold Cross Validation adalah 68,6%, 67,63%, 71,52% dan 69,34%.

(Efrizoni et al., 2022) Hasil dari percobaan menunjukkan bahwa algoritma Na¨ive Bayes memiliki akurasi tertinggi dengan menggunakan ekstraksi fitur TF-IDF sebesar 87% dan BoW sebesar 83%. Untuk ekstraksi fitur Doc2Vec, akurasi tertinggi pada algoritma SVM sebesar 81%. Sedangkan ekstraksi fitur Word2Vec dengan algoritma machine learning (i.e. i.e. Na¨ive Bayes, Support Vector Machines, Decision Tree, K-Nearest Neighbors, Random Forest, Logistic Regression) memiliki akurasi model dibawah 50%. Hal ini menyatakan, bahwa Word2Vec kurang optimal digunakan bersama algoritma machine learning, khususnya pada dataset tribunnews.com.

### 2.2.1. Matriks Penelitian

**Tabel 2.4** Matrix penelitian

| No | Judul | Nama dan tahun penelitian | Klasifikasi nama | Machine learning | Fitur ekstraksi | | Model Klasifikasi | | Dataset indonesia | Evaluasi model | | Implementasi ke dalam aplikasi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Bag-of-Word | N-grams | Logistic Regressio | Multinomial Naive Bayes | | Akurasi | *F1-score* | |
| 1 | Predicting the Gender of Indonesian Names | (Septiandri, 2017) | ✓ | ✓ | - | ✓ | ✓ | - | ✓ | ✓ | ✓ | - |
| 2 | Gender Classification of Thai Facebook Usernames | (Yuenyong & Sinthupinyo, 2020) | ✓ | ✓ | - | - | - | ✓ | - | ✓ | - | - |
| 3 | A Computational Linguistic Approach for Gender Prediction Based on Vietnamese Names | (Ho Huong et al., 2022) | ✓ | ✓ | ✓ | ✓ | ✓ | - | - | ✓ | - | - |
| 4 | Customer gender prediction system on hierarchical E-commerce data | (Khan et al., 2020) | ✓ | ✓ | ✓ | ✓ | ✓ | - | - | ✓ | ✓ | - |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | Predicting Gender of Brazilian Names Using Deep Learning | (Rego et al., 2021) | ✓ | - | - | ✓ | - | - | - | ✓ | ✓ | |
| 7 | Advance gender prediction tool of first names and its use in analysing gender disparity in Computer Science in the UK, Malaysia and China | (Zhao & Kamareddine, 2018) | ✓ | - | - | - | - | - | - | ✓ | ✓ | ✓ |
| 8 | Klasifikasi Jenis Kelamin Pengguna Twitter dengan menggunakan Metode BM25 dan K-Nearest Neighbor (KNN) | (Zakia, 2020) | ✓ | ✓ | - | - | - | - | ✓ | ✓ | ✓ | - |
| 9 | Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning | (Efrizoni et al., 2022) | - | ✓ | ✓ | ✓ | ✓ | - | - | ✓ | ✓ | - |