BAB 2
TINJAUAN PUSTAKA

2.1 Review Literatur

Dalam penyusunan skripsi ini, penulis mereferensi penelitian-penelitian terdahulu yang terkait dengan topik latar belakang masalah yang sejenis, sehingga dapat menjadi pendukung untuk proses pengimplementasian sesuai latar belakang masalah skripsi. Adapun beberapa riset tersebut adalah

Tabel 2. 1 Review Literatur

REF	DATA SET	METODE	KLASIFIKASI	KELEBIHAN	KEKURANGAN
[1]	Data berita sebanyak 500 terdiri dari 143 Hoaks dan 357 berita fakta	Seleksi fitur berdasarkan Mutual Information dan Information Gain	Naïve Bayes	nilai yang dihasilkan oleh metode NBC dengan tambahan seleksi fitur MI dan IG lebih besar dibandingan dengan metode NBC yang tidak menggunakan seleksi fitur dalam mengidentifikasi berita hoaks	Disarankan menggunakan tambahan metode lexical chain disertai metode word sense disambiguation yang berguna untuk mengidentifikasi makna yang ada di dalam sebuah data, sehingga pengidentfikasian hoaks dapat dilakukan lebih baik

[2]	dataset berisi 9727 berita berlabel hoax dan 474 berita berlabel valid	chi-square dan information gain	Naive Bayes Classifier	Berdasarkan hasil stratified k-fold cross validation diperoleh nilai akurasi tertinggi yaitu 0,87198 untuk klasifikasi menggunakan seleksi fitur chi-square. Nilai akurasi terendah didapatkan pada klasifikasi menggunakan seleksi fitur information gain dengan nilai akurasi sebesar 0,83541.	Perbandingan yang jauh antara data berita dan hoaks

[3]	Data berjumlah 600 Data yang tergolong fakta berjumlah 372 artikel berita. Data yang tergolong hoaks berjumlah 228 artikel berita	Maximum Entropy dengan Seleksi Fitur Information Gain	Algoritma Maximum Entopy	Penelitian ini menghasilkan akurasi tertinggi sebesar 0,8 dengan seleksi fitur information gain (threshold = 50%)	Pada algoritme Maximum Entropy, dokumen yang berisi fitur kata dengan kemunculan kata yang tinggi pada satu kelas sangat memengaruhi hasil klasifikasinya.

[4]	sejumlah 3336 catatan berita diambil dari 5 penyedia berita	Sequential Forward Selection, Sequential Backward Elimination, Information Gain, Inverse Document Frequency, Term Frequency	SVM, Naive Bayes, Decision Trees	metode yang diusulkan menunjukkan hasil yang lebih baik dibandingkan metode yang ada. Metode klasifikasi dapat diubah tergantung pada distribusi dataset yang diberikan	hasil menunjukkan bahwa metode seleksi fitur memberikan hasil terbaik untuk sebagian besar situasi, terlepas dari distribusi data. Akurasi dapat ditingkatkan dengan menggunakan teknik bagging dan boosting.

[5]	400 data dengan jumlah 200 data fakta dan 200 data hoaks	Seleksi gain ratio	K-Nearest Neighbor	penggunaan seleksi fitur Gain Ratio dengan Information Gain rata-rata lebih stabil nilai pada f-measure yang menggunaka n seleksi fitur Gain Ratio di setiap nilai k yang digunakan karena terdapat perbedaan pada pemilihan term yang berdasarkan nilai threshold.	saran untuk penelitian selanjutnya adalah dengan menerapkan metode lain selain KNN sehingga bisa menjadi pembanding untuk penggunaan seleksi fitur Gain Ratio. Menambahkan kelas netral pada data juga dapat dilakukan untuk melihat hasil dari penggunaan Gain Ratio.

Penelitian yang dilakukan adalah mengidentifikasi berita atau informasi Hoaks dalam bentuk teks. Tujuannya untuk mengetahui tingkat akurasi dari kombinasi metode tersebut dalam mengklasifikasi berita Hoaks berupa teks. Dengan dataset berita Hoaks dan berita fakta yang sudah tersedia maka kombinasi metode ini diharapkan dapat memperoleh hasil yang lebih baik dibandingkan dengan penelitian sebelumnya yang pernah dilakukan.

2.2 Landasan Teori

2.2.1 Berita Dan Hoaks

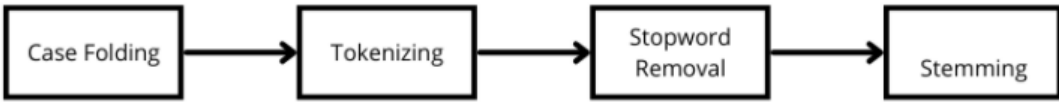
Berita adalah narasi atau deskripsi tentang peristiwa atau kejadian yang aktual, berupa laporan atau pengumuman. Menurut Kamus Besar Bahasa Indonesia, berita dapat diartikan sebagai pemberitahuan tentang suatu kejadian yang sedang ramai diperbincangkan. Berita merupakan sebuah narasi yang bertujuan untuk menarik perhatian, memberikan informasi, atau memberikan pembelajaran kepada pembaca. Secara esensial, berita adalah catatan tentang aktivitas manusia yang belum dipublikasikan, yang berusaha untuk memberikan gambaran yang jelas tentang suatu peristiwa atau kejadian yang terjadi pada waktu tertentu[8].

Hoaks adalah informasi palsu atau manipulatif yang disebarkan dengan sengaja untuk menyesatkan atau menipu orang lain. Biasanya, hoaks berupa berita palsu, foto yang diedit, atau video manipulatif yang tidak memiliki dasar fakta yang valid. Penyebaran hoaks dapat dilakukan melalui berbagai media, termasuk media sosial, situs web, atau aplikasi pesan instan. Dampaknya dapat merugikan, menciptakan ketidakpercayaan terhadap informasi yang sebenarnya valid, serta memicu kekacauan sosial atau politik.

Berita hoaks dibuat dengan tujuan untuk mempengaruhi pembaca agar melakukan tindakan yang bertentangan dengan kebenaran atau mencegah mereka melakukan tindakan yang benar. Biasanya, metode yang digunakan adalah ancaman, penyesatan, atau membuat pembaca mempercayai hal-hal yang tidak benar. Hoaks dapat diciptakan dengan berbagai cara, termasuk penggunaan data pribadi yang diperoleh dengan menyesatkan korban bahwa data tersebut diperlukan untuk keperluan resmi.

2.2.2 *Preprocessing*

Preprocessing teks merupakan langkah penting dalam mengubah data yang memiliki struktur yang tidak teratur menjadi data yang terstruktur, sehingga memungkinkan untuk dilakukan analisis dan pemrosesan lebih lanjut sesuai dengan kebutuhan. Data yang telah diproses secara tepat akan memfasilitasi pemahaman dan ekstraksi informasi yang lebih baik dari teks yang tidak terstruktur, seperti dokumen teks media sosial[7]. *Preprocessing* terdiri dari beberapa langkah pada gambar berikut.



Gambar 2. 1 *Preprocessing*

Langkah-langkah dalam preprocessing teks meliputi beberapa tahapan, di antaranya adalah *case folding, tokenizing, Stopword Removal, stemming*, dan penghitungan bobot kata. *Case folding* dilakukan untuk menyamakan representasi huruf ke dalam bentuk yang konsisten, misalnya dengan mengonversi semua huruf menjadi huruf kecil. *Tokenizing* adalah proses memecah kalimat atau teks menjadi token-token kata yang lebih kecil untuk mempermudah analisis. *Stopword Removal* dilakukan untuk menghilangkan kata-kata yang tidak relevan atau tidak informatif, seperti stop words. Selanjutnya, *stemming* digunakan untuk mengubah kata-kata menjadi bentuk dasarnya agar variasi kata yang sama dapat dihitung sebagai satu entitas. Terakhir, penghitungan bobot kata dilakukan untuk memberikan nilai numerik pada setiap kata berdasarkan frekuensi kemunculannya dalam dokumen, sehingga memungkinkan analisis lebih lanjut menggunakan algoritma seperti *Naïve Bayes*. Langkah-langkah ini merupakan bagian integral dalam mengoptimalkan kinerja model *Naïve Bayes* dengan mempersiapkan data teks secara efisien sebelum dilakukan proses klasifikasi [6].

2.2.3 *Case Folding*

Case folding adalah salah satu langkah penting dalam pra-pemrosesan teks yang bertujuan untuk mengubah semua huruf dalam teks menjadi huruf kecil atau besar. Tujuannya adalah untuk menyamakan representasi teks yang berbeda, Langkah ini membantu mengurangi kompleksitas dan mempermudah analisis teks. tahap awal dari preprocessing text in mengubah karakter huruf teks menjadi huruf kecil semua. Karakter yang diterima hanya 'a' hingga 'z' [6]. Pada proses ini juga dilakukan penghilangan angka, tanda baca dan karakter lain yang dianggap tidak memiliki pengaruh terhadap pemrosesan teks.

2.2.4 *Tokenizing*

Tokenizing adalah proses memecah teks menjadi unit-unit diskrit yang disebut token. Token dapat berupa kata, frasa, atau karakter terpisah, tergantung pada kebutuhan analisis. Proses *tokenizing* memungkinkan teks untuk diuraikan menjadi unit-unit yang lebih kecil sehingga dapat diproses lebih lanjut, seperti menghitung frekuensi kemunculan kata atau melakukan analisis sintaksis dan tahap pemotongan string input berdasarkan tiap kata yang menyusunnya [6].

2.2.5 *Stopword Removal*

Stopword Removal adalah langkah penting dalam pemrosesan dokumen yang bertujuan untuk menentukan kata-kata yang akan digunakan untuk merepresentasikan dokumen. Selain memvisualisasikan isi dokumen, kata-kata ini juga berfungsi untuk membedakan satu dokumen dengan yang lain dalam kumpulan dokumen. Proses ini melibatkan penghapusan *stop words*, yaitu kata-kata yang tidak memberikan kontribusi signifikan dalam pemahaman konten dokumen. Penghapusan *stopwords* diperlukan untuk memfokuskan analisis pada kata-kata yang lebih relevan dan deskriptif. Selain itu, langkah ini membantu meningkatkan akurasi analisis teks dengan meminimalkan gangguan dari kata-kata yang tidak relevan. Dalam bahasa Indonesia, contoh stop words seperti "yang", "dan", "dari", "di", "seperti" dan lainnya [6].

2.2.6 *Stemming*

Stemming merupakan langkah lanjutan dalam pemrosesan teks yang bertujuan untuk mencari akar kata dari kata-kata yang telah melalui proses *Stopword Removal*. Pada tahap ini, dilakukan identifikasi dan pengambilan berbagai bentuk kata ke dalam representasi yang seragam. *Stem*, atau akar kata, merupakan bagian dari kata yang tetap setelah dihilangkan imbuhannya, termasuk awalan dan akhiran. Sebagai contoh, kata "beri" adalah stem dari kata-kata seperti "memberi", "diberikan", "memberikan", dan "pemberian".

Proses stemming sangat penting dalam analisis teks karena membantu dalam mengelompokkan kata-kata dengan makna yang sama, sehingga mempermudah dalam pemahaman dan klasifikasi teks yang lebih efisien. Stemming juga membantu mengurangi kompleksitas data dengan mengonversi variasi kata-kata ke dalam bentuk yang lebih sederhana, sehingga memfasilitasi proses analisis lebih lanjut [6].

2.2.7 *Vectorizing*

Agar dapat membangun model dari dataset yang berisi kalimat atau kata-kata, diperlukan suatu proses untuk mengubah dataset yang awalnya berupa string menjadi representasi numerik yang dapat dipahami oleh mesin. Transformasi ini memungkinkan mesin untuk memproses dan menganalisis makna dari setiap baris data secara lebih efektif.

Count Vectorizer adalah teknik pemrosesan teks dalam *Natural Language Processing* (NLP) yang digunakan untuk mengubah teks menjadi representasi numerik atau vektor fitur. *CountVectorizer*, berfungsi untuk menghitung frekuensi kata dalam dokumen. *Count Vectorizer* dapat mengubah fitur teks menjadi sebuah representasi vector.

2.2.8 Seleksi Fitur Gain Ratio

Dalam konteks klasifikasi, *gain ratio* memiliki peranan penting dalam mengevaluasi kebermaknaan suatu fitur dalam mengidentifikasi kelas tertentu, dengan membandingkan frekuensi kemunculan fitur tersebut di seluruh kelas. Dengan demikian, *gain ratio* menjadi alat yang berguna dalam mengukur kontribusi relatif dari setiap fitur terhadap proses klasifikasi secara keseluruhan.

Dalam analisis teks, *Gain Ratio* dapat digunakan untuk mengevaluasi pengaruh leksikal suatu kata terhadap kelas-kelas yang mungkin ada dalam dataset. Proses ini melibatkan perhitungan jumlah kata yang digunakan dalam setiap kelas dan kemudian menentukan seberapa besar informasi yang diberikan oleh masing-masing kata terhadap pemisahan kelas tersebut.

Perhitungan *gain ratio* dimulai dari perhitungan *mutual information* seperti. *Gain ratio* adalah pembagian antara *mutual information* dan *entropy*. *Mutual information* merupakan nilai ukur yang menyatakan keterikatan atau ketergantungan antara dua variabel atau lebih. Berikut adalah formual untuk menghitung *mutual information*[3]*:*

$$MI(X,Y) = \sum_y \sum_x P(x,y)\ \log_2 \frac{P(x,y)}{P(x)P(y)} \qquad \text{Persamaan 2. 1}$$

Keterangan

$MI(X,Y)$	: *mutual information fitur X dan Y*
X	: kolom fitur
Y	: kolom label
Σ	: total penjumlahan setiap variabel
$P(x,y)$	: joint probabilitas x dan y(probabilitas di mana x dan y muncul secara bersamaan)
$P(x)$	: probabilitas x
$P(y)$	: probabilitas y
x	: *Value* dari kolom X
y	: *Value* dari kolom Y

Setelah mendapatkan nilai *mutual information*, maka langkah selanjutnya adalah menghitung *entropy*. Entropy digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan keluaran atribut. Berikut adalah formula untuk menghitung *entropy* [6].

$$E(Y) = \sum_{y} P(y) \log_2 \frac{1}{P(y)} \qquad \text{Persamaan 2. 2}$$

Keterangan

$E(Y)$: *Entropi target class* Y (Label)

Σ : total penjumlahan setiap variabel

$P(x)$: probabilitas x

Y : kolom Label

y : *Value* dari kolom Y

Maka dari itu penghitungan *Gain Ratio* adalah hasil dari penghitungan *Mutual Information* dibagi dengan hasil penghitungan *Entropy*. Berikut adalah formula akhir untuk menghitung *Gain Ratio* [6] [7]*:*

$$Gain\ Ratio(i) = \frac{MI(X,Y)}{E(X)} = \frac{\sum_y \sum_x P(x,y)\ \log_2 \frac{P(x,y)}{P(x)P(y)}}{\sum_x P(x) \log_2 \frac{1}{P(x)}} \qquad \text{Persamaan 2. 3}$$

Keterangan

Σ : total penjumlahan setiap variabel

$P(x,y)$: joint probabilitas x dan y(probabilitas di mana x dan y muncul secara bersamaan)

$P(x)$: probabilitas x

$P(y)$: probabilitas y

X : kolom fitur

Y : kolom label

x : *Value* dari kolom X

y : *Value* dari kolom Y

2.2.9 Pemilihan Fitur

Pemilihan fitur, atau yang sering disebut sebagai *feature selection* merupakan sebuah langkah dalam proses pengembangan model. Ini melibatkan identifikasi dan pemilihan subset fitur atau atribut yang dianggap paling relevan dan informatif untuk dimasukkan ke dalam model atau untuk analisis lebih lanjut. Langkah ini dilakukan dengan tujuan untuk meningkatkan kinerja model secara keseluruhan atau untuk mengurangi kompleksitas model dengan menghilangkan fitur-fitur yang tidak memberikan kontribusi signifikan atau bahkan redundan dalam pembentukan model [7].

2.2.10 Naive Bayes

Naive bayes adalah metode yang digunakan dalam statistika untuk menghitung peluang dari suatu hipotesis, *Naïve Bayes* menghitung peluang suatu kelas berdasarkan pada atribut yang dimiliki dan menentukan kelas yang memiliki probabilitas paling tinggi. *Naive Bayes* mengklasifikasikan kelas berdasarkan pada probabilitas sederhana dengan mangasumsikan bahwa setiap atribut dalam data tersebut bersifat saling terpisah. Metode *Naive Bayes* merupakan salah satu metode yang banyak digunakan berdasarkan beberapa sifatnya yang sederhana, metode *Naive Bayes* mengklasifikasikan data berdasarkan probabilitas P atribut x dari setiap kelas y data. Pada model probablitas setiap kelas k dan jumlah atribut a yang dapat dituliskan seperti Persamaan berikut [6]:

$$p(y_k|x_1, x_2, x_3, \ldots, x_a) \qquad \text{Persamaan 2. 4}$$

Penghitungan *Naïve Bayes* yaitu probabilitas dari kemunculan dokumen xa pada kategori kelas yk P(xa|yk), dikali dengan probabilitas kategori kelas P(yk). Dari hasil kali tersebut kemudian dilakukan pembagian terhadap probabilitas kemunculan dokumen P(xa). Sehingga didapatkan rumus penghitungan Naïve Bayes dituliskan pada Persamaan berikut [6]

$$P(y_k|x_a) = \frac{P(y_k)P(x_a|y_k)}{P(x_a)}$$ Persamaan 2. 5

Keterangan

$P(y_k|x_a)$: probabilitas y_k terhadap x_a

Σ : total penjumlahan setiap variabel

$P(y_k)$: probabilitas y_k

$P(x_a)$: probabilitas x_a

2.2.11 Laplace Smoothing

Dalam klasifikasi *Naive Bayes*, saat menghitung probabilitas $P(d|C_i)$ untuk sebuah atribut d yang akan diklasifikasikan, diasumsikan bahwa atribut-atribut tersebut bersyarat independen dan mencari nilai probabilitas tersebut. Namun, terkadang mendapatkan nilai probabilitas nol untuk beberapa kasus atau bahkan tidak terdefinisi. Hal ini menyebabkan probabilitas juga menjadi nol. Untuk menghindari hal ini, kita mengasumsikan bahwa basis data pelatihan D sangat besar sehingga penambahan satu ke setiap penghitungan yang kita butuhkan hanya akan membuat perbedaan yang tidak signifikan dalam estimasi nilai probabilitas, sehingga menghindari kasus nilai probabilitas nol. Teknik ini untuk estimasi probabilitas dikenal sebagai koreksi Laplace atau estimator Laplace.

Dalam koreksi Laplace menghasilkan perkiraan probabilitas yang lebih akurat. Jika kita memiliki 'q' penghitungan yang akan ditambahkan satu, maka harus menambahkan q ke penyebut yang sesuai yang digunakan dalam perhitungan probabilitas sesuai dengan jumlah class. Teknik ini membantu membuat prediksi lebih akurat dengan menghindari probabilitas nol dan memperbaiki ketidaksempurnaan yang mungkin muncul dalam estimasi probabilitas berdasarkan dataset pelatihan yang terbatas. Dengan demikian, koreksi *Laplace* memainkan peran penting dalam meningkatkan kualitas prediksi dari model klasifikasi *Naive Bayes* [8].

$$P(t) = \frac{number\ of\ time\ t\ occurs + 1}{number\ of\ possible\ outcomes\ + number\ of\ unique\ class}$$ Persamaan 2. 6

2.2.12 Confusion Matrix

Confusion Matrix adalah tahap analisis dan evaluasi terhadap performansi sistem yang dirancang. Performansi diukur dengan nilai *precision*, dan *recall*. Confusion matrix adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Tabel *Confusion matrix* dapat dilihat dalam Tabel berikut

Tabel 2. 2 Confusion *Matrix* [5]

	Prediksi Kelas Ya (Hoaks)	**Prediksi Kelas Tidak (tidak hoaks)**
Aktual kelas Ya (Hoaks)	*True Positive %* (TP)	. *False Positive %* (FP)
Aktual Kelas Tidak (tidak Hoaks)	*False Negative %* (FN)	*True Negative %* (TN)

True Positive (TP) merupakan kelas hasil klasifikasi dan kelas sesungguhnya sama-sama Hoaks. *False Positive* (FP) merupakan kelas hasil klasifikasi Hoaks dan kelas sesungguhnya Tidak hoaks. *True Negative* (TN) merupakan kelas hasil klasifikasi dan kelas sesungguhnya sama-sama Tidak Hoaks. *False Negative* (FN) merupakan kelas hasil klasifikasi Tidak Hoaks dan kelas sesungguhnya *x*.

Pada tahapan ini akan diukur nilai performansi dari model yang telah dibuat melalui proses perhitungan *precision, recall, F1-Score*. Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Sedangkan recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual.

1. *Precision*

Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Dalam *data mining*, *precision* adalah jumlah dokumen yang dengan benar diklasifikasikan dalam sebuah kelas dibagi jumlah total dokumen dalam kelas tersebut secara sederhana *precision* menunjukkan nilai benar positif dibagi seluruh nilai yang diperkirakan positif. Dengan persamaan [5]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Persamaan 2. 7}$$

Keterangan

TP : *True Positive*

FP : *False Positive*

2. *Recall*

Recall adalah jumlah pengguna yang dengan benar diklasifikasikan dalam sebuah kelas dibagi dengan jumlah total pengguna dalam kelas tersebut . *Recall* menunjukkan perbandingan antara nilai benar positif dengan seluruh data yang sebenarnya Berikut rumus dari recall [5]:

$$Recall = \frac{TP}{TP + FN} \qquad \text{Persamaan 2. 8}$$

Keterangan

TP : *True Positive*

FN : *False Negative*

3. *F1 Score*

F1 Score merupakan metrik yang biasa digunakan untuk mengevaluasi kinerja model klasifikasi biner Skor F1 berkisar antara 0 hingga 1, dengan skor 1 menunjukkan presisi dan perolehan sempurna, serta 0 menunjukkan kinerja buruk. *F1 Score* yang tinggi berarti model tersebut memiliki presisi dan recall yang tinggi,

yang menunjukkan bahwa model tersebut merupakan model yang baik untuk tugas klasifikasi biner . *F1 Score* adalah nilai rata-rata dari perbandingan *precision* dan *recall*. Berikut rumus dari F1 dapat dilihat sebagai berikut[9]:

$$F1\ Score\ = 2 * \frac{Recall * Precission}{Recall + Precission}$$ Persamaan 2. 9