Analisis Algoritma Naïve Bayes Classifer Untuk Mendeteksi Berita Hoax Pada Dinas Kominikasi Informatika Dan Persandian

Anjasmara Tarigan[1], Ilham Sahputra[2], Teuku Multazam[3]

[1,2] Program Studi Sistem Informasi, Universitas Malikussaleh
[3]Program Studi Teknik Elektro, Universitas Malikussaleh
Jln. Kampus Unimal Bukit Indah, Blang Pulo, Kec. Muara Satu,
Kabupaten Aceh Utara, Aceh, 24355
Corresposnding Author: anjasmara.200180053@mhs.unimal.ac.id

Abstrak

Masyarakat mampu mengkonsumsi tiap informasi yang tersebar di internet dengan cepat dan terkadang informasi yang beredar tidak selalu memberikan kebenaran yang sesuai dengan kenyataannya (hoax). Demi mendapatkan keuntungan dan mencapai tujuan pribadi, hoax seringkali sengaja dibuat dan dibagikan. Informasi yang didapatkan dari hoax tentunya dapat mempengaruhi masyarakat karena menimbulkan keraguan dan kebingungan terhadap informasi yang diterima. Oleh karena itu, penulis membahas tentang bagaimana mengklasifikasikan berita hoax menggunakan algoritma Naïve Bayes Classifier sehingga mampu memprediksi sebuah berita hoax atau fakta. Dataset yang dikumpulkan sebanyak 200 diantaranya 100 fakta dan 100 hoax. Dalam proses data split dengan pembagian data testing dan data trainning dengan pembagian 60% data training dan 40% data testing yaitu 120 data trainning dan 80 data testing. Dengan menggunakan algoritma Naive Bayes Classifer mendapatkan nilai Akurasi Sebesar 97%. Pada penelitian ini juga membahas tentang pemanfaatan TF-IDF dalam melakukan Klasifikasi Berita Hoax dengan menggunakan Algoritma Naive Bayes Classifer. Dengan nilai Akurasi 97% menyatakan bahwa Algoritma Naive Bayes Classifer efektif dalam melakukan klasifikasi. Algoritma in mengandalkan teorema Bayes untuk menghitung probabilitas berita hoax berdasarkan kemunculan kata-kata atau fitur-fitur lainnya dalam berita.

Kata Kunci : Naïve Bayes Classifier, Klasifikasi Berita Hoax, Berita Hoax

Abstract

The public is able to consume any information that is spread on the internet quickly and sometimes the information that is circulating does not always provide the truth according to reality (hoax). In order to gain profit and achieve personal goals, hoaxes are often deliberately created and shared. Information obtained from hoaxes can certainly affect the public because it creates doubts and confusion about the information received. Therefore, the author discusses how to classify hoax news using the Naïve Bayes Classifier algorithm so that it can predict hoax news or facts. There are 200 datasets collected, including 100 facts and 100 hoaxes. In the data split process by dividing testing data and trainingdata by dividing 60% of training data and 40% of testing data, namely 120 training data and 80 testing data. By using the Naive Bayes Classifier Algorithmto get an accuracy value of 97%. This study also discusses the use of TF-IDF in classifying hoax news using the Naive Bayes Classifier Algorithm. With an accuracy value of 97%, it states that the Naive Bayes Classifier Algorithm is effective in classifying. This algorithm relies on Bayes' theorem to calculate the probability of hoax news based on the appearance of words or other features in the news.

Keywords: Naïve Bayes Classifier, Support Vector Machine, Hoax NewsClassification, Hoax News

1. PENDAHULAN

Persebaran arus informasi melalui internet saat ini sangatlah mudah dan cepat dimana waktu serta jarak tidak menjadi sebuah penghalang. Jumlah pengguna internet di Asia dimana pada Maret 2021 Indonesia telah mencapai 212,35 juta jiwa pengguna internet. Menggunakan data tersebut, Indonesia menempati peringkat ketiga diantara Negara-negara dengan pegguna internet terbanyak di Asia (Kusnandar, 2021) . Sehingga dengan meningkatnya pengguna internet, masyarakat dapat mengkonsumsi tiap informasi yang tersebar dengan cepat. Dengan kecepatan tersebut tentu saja menghasilkan dampak positif maupunnegative, dimana informasi yang beredar tidak selalu memberikan kebenaran yang sesuai dengan kenyataan atau bisa disebut dengan hoax.

Hoax adalah informasi atau berita yang mengandung hal-hal yang belum teridentifikasi atau bukan fakta yang sebenarnya terjadi (Juditha, 2018). Demi mendapatkan keuntungan dan mencapai tujuan pribadi, hoax sering kali sengaja dibuat dan dibagikan sehingga dapat menyebar lebih cepat. Informasi yang didapatkan dari hoax tentunya dapat mempengaruhi masyarakat karena menimbulkan keraguan dan kebingungan terhadap informasi yang di terima,serta mampu merusak citra individu dan kelompok yang berkaitan. Setidaknya 30% hingga hampir 60% masyarakat Indonesia terpapar hoax saat mengakses danberkomunikasi melalui dunia maya. Sementara hanya 21% hongga 36% saja yang dapat mengenali atau mendeteksi adanya hoax. Sebagian hoax yang ditemukan termasuk isu politik,kesehatan,dan agama (Cahyadi, 2020).

Solusi saat ini dari pihak kominfo telah menyediakan situs kominfo.go.id layanan pengaduan konten dari berbagai sumber media social yang diduga berisi berita hoax. Melalui layanan pengaduan konten memang memberikan fasilitas bagi masyarakat, apabila menemukan adanya website/situs, konten media social, game online yang melanggar aturan perundangan di Indonesia. Penyebaran berita hoax memiliki dampak yang sangat negative terhadap individu dan masyarakat. Oleh karena itu untuk mendeteksi suatu informasi atau berita yang berisikan hoax,diperlukan adanya alat bantu yang dapat mendeteksi kebenaran berita tersebut apakah termasuk hoax atau valid. Metode klasifikasi diperlukan sebagai salah satu cara untuk mengurangi tersebarnya konten berisi hoax dengan memanfaatkan text mining dan menggunakan python sebagai bahasa pemrogramannya.

Text mining merupakan proses yang melibatkan transformasi teks tidak terstruktur menjadi format terstruktur yang dapat dianalisis oleh komputer. Tujuandari proses ini adalah untuk mengidentifikasi pola atau informasi baru yang berguna dalam konteks text mining dan machine learning. Algoritma yang digunakan dalam kasus ini adalah Naïve Bayes Classifier. Algoritma ini dipilih karena memiliki keunggulan dalam kemudahan pemahaman, memberikan hasil yang baik, dan efektif dalam membangun model untuk keperluan analisis teks. Dengan menggunakan Naïve Bayes Classifier, kita dapat mengolah data teks secara efisien dan mendapatkan wawasan yang berharga dari data tersebut.

2. TINJAUAN PUSTAKA

2.1 Berita Hoax

Dalam praktek jurnalistik, berita menduduki posisi utama dan menurut pakar jurnalistik untuk mendefenisikan berita itu sangatlah sulit. Belum ada batasan yang dapat mencakup seluruh segi, sifat, dan karakter, ciri dan jenis- jenisnya. Berita adalah segala laporan mengenai peristiwa, kejadian, gagasan,fakta, yang menarik perhatian dan penting untuk disampaikan atau dimuat dalam media massa agar diketahui atau menjadi kesadaran umum.

Berita sebenarnya berasal dari bahasa sansekerta, yaitu Vrit yang dapat dimaknai dengan Vritta dalam bahasa Inggris, memiliki arti „ada‟ atau „terjadi‟. Beberapa orang memaknainya dengan Vritta, yang berarti "kejadian" atau sebuah peristiwa yang telah terjadi‟. Dalam bahasa Indonesia Vritta memiliki artiyaitu sebuah „berita atau warta‟. Sedangkan menurut KBBI, berita merupakan cerita atau keterangan mengenai kejadian atau peristiwa yang hangat (ALVINA, 2023).

Hoax merupakan manipulasi berita yang sengaja dilakukan dan bertujuan untuk memberikan pengakuan dan pemahaman yang salah. Hal itu sebenarnya sudah terjadi sejak lama, namun kecanggihan teknologi membuat penyebaran kabar tersebut menjadi lebih luas dan menjadi prestasi tersendiri bagi sang pembuat hoax jika berhasil menyebarluaskannya (Alfarisi, 2023).

2.2 Machine Learning

Machine learning atau pembelajaran mesin adalah pendekatan dalam Artificial Intelligence (AI) yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi. Machine Learning mencoba menirukan bagaimana proses manusia atau makhluk cerdas belajar dan mengeneralisasi. Dua aplikasi utama dalam machine mearning yaitu, klasifikasi dan prediksi. Ciri khas dari machine learning adalah adanya proses pelatihan, pembelajaran, atau training. Oleh karena itu, machine mearning membutuhkan data untuk dipelajari yang disebut sebagai data training. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan dalam text mining adalah sekumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari text mining antara lain yaitu pengkategorisasian teks dan pengelompokkan teks (Baby, 2019).

2.3 Text Mining

Text mining memiliki cakupan yang luas dan umumnya bertujuan untuk mengurai data tekstual untuk mengektrak fakta yang dapat dibaca oleh prangkat elektronik (Mirzaalian, 2019). Menurut (Luqyana, 2018)Text mining merupakan ilmu yang bertujuan untuk memproses teks agar menjadi informasi, menambang suatu data yang berupa teks yang bersumber dari data terebut. Data biasanya diproleh dari dokumen dan digunakan untuk mencari kata-kata yang dapat mewakili isi dari dokumen tersebut (Farach, 2019).

2.4 Text Preprocessing

Text proprecessing merupakan tahapan awal pada Text Mining yang merupakan suatu proses yang mempersiapkan teks agar dapat diolah lebih lanjut. Preprocessing secara umum bertujuan untuk mengubah informasi dari tiap-tiap sumber data ke dalam bentuk atau format

yang baku sebelum menerapkan berbagai metode-metode pengambilan data terhadap dokumen yang akan diproses (Fitria, 2018).

2.5 TF-IDF

Metode TF-IDF merupakan perhitungan yang mendeskripsikan seberapa pentingnya sebuah kata (term) terhadap sebuah dokumen dengan memberikan bobot pada setiap kata. Frekuensi kata adalah ukuran seringnya kemunculan kata dalam sebuah teks dan juga pada seluruh teks dalam korpus. TF-IDF digunakan untuk mengubah teks menjadi angka yang dapat diproses pada Machine Learning, lalu dihitung jumlah kemunculan katanya dalam sebuah teks (Munjiah Nur Saadah, 2019). Menggunakan TF-IDF untuk pemberian bobot term pada dataset atau dokumen menggunakan urutan token berupa unigram dalam implementasinya sehingga jumlah token dari TF-IDF hanya satu kata saja. Berikut merupakan rumus untuk perhitungan TF-IDF:

$$F = \left\{\frac{1+ \log_{10}(f_{t,d})}{0}\right\} \quad (2.1)$$

$$IDF_j = \log \frac{D}{df_j} \quad (2.2)$$

$$w_{ij} = tf_{ij} x \log\left(\frac{D}{df_j}\right) + 1 \quad (2.3)$$

Keterangan:

$ft,$ = frekuensi term (t) pada dokumen (d)

D = jumlah semua dokumen dalam dataset

dfj = jumlah dokumen yang mengandung term(tj)

wij = bobot term (tj) terhadap dokumen (d)

$tfij$ = jumlah kemunculan term (tj) dalam dokumen (d)

2.6 Naïve Bayes Classifer (NBC)

Algoritma NBC merupakan metode klasifikasi yang statistik berdasarkan pada teorema Bayes (Baby, 2019). NBC merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan Naive dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi Naive Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. NBC berpotensi baik untuk mengklasifikasikan data karena kesederhanaannya (ing, 2020). Persamaan NBC untuk klasifikasi terdapat pada Persamaan berikut.

$$p(W_i|W_j) = \frac{N_{cw}+1}{N_c+1} \quad (2.4)$$

Dimana:

a) Ncw = Jumlah kata wi yang ada dalam dokumen training yang masuk ke dalam kategori
b) Cj. Nc = Jumlah semua kata yang ada dalam dokumen training yang masuk kedalam kategori Cj (tanpa menghiraukan ada kata ang sama atau tidak).
c) V = Jumlah total jenis kata yang ada dalam dokumen training (kata yang sama hanya dihitung 1).

2.7 Klasifikasi

Klasifikasi adalah kegiatan atau pekerjaan mengelompokkan data atau mempelajari sesuatu menurut ciri-ciri yang dimilikinya. Ini dapat dilakukan agar setiap objek diberi kategori tertentu. Klasifikasi adalah pekerjaan yang melakukan pembuatan sebuah model dengan dasar data latih yang telah disiapkan, selanjutnya dengan model tersebut akan dilakukan klasifikasi data yang baru, tujuannya adalah agar sistem yang dibuat akan dapat melakukan pengklasifikasian data terhadap semua data set dengan baik dan benar, selanjutnya keberhasilan hasilnya akan diukur setelah polanya terbentuk (Widiastuti, 2023).

2.8 Confusion Matrix

Confusion Matrix memberikan informasi yang dapat digunakan untuk merangkum kinerja dari klasifikasi yang sehubungan dengan beberapa test data dengan memvisualisasikan. Tabel 3.1 menunjukan matrix dua dimensi, dimana satu dimensi diindeks oleh kelas sebenarnya dari suatu objek dan di dimensi lain oleh kelas yang ditetapkan oleh classifier (Kohli, 2019).

		Kelas Prediksi	
		Negative(N)	*Positif* (P)
Kelas Aktual	*Negative*	*True Negative* (TN)	*False Positive*(FP)
	Positive	*False Negative*(FN)	*True Positive*(TP)

Gambar 1. Confusion Matrix

Keterangan:

a. True Positive (TP) merupakan hasil prediksi yang dikeluarkan program berupa positif dan itu memang benar (positif).
b. True Negative (TN) merupakan hasil prediksi yang dikeluarkan program berupa negatif dan itu memang benar (negatif).
c. False Positive (FP) merupakan hasil prediksi yang dikeluarkan program berupa positif, namun hasil aktualnya yaitu negatif.
d. False Negative (FN) merupakan hasil prediksi yang dikeluarkan program berupa negatif, namun hasil aktualnya yaitu positif.

x merupakan visualisasi untuk mengukur kualitas performa prediksi dari algoritma klasifikasi dengan menampilkan precision, recall, f1-score, support dan accuracy (Kohli, 2019)

Precision atau presisi digunakan untuk menghitung berapa rasio / keakuratan dari prediksi yang dilakukan oleh model itu benar.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2.5)$$

Recall merupakan kemampuan classifier untuk menemukan semua kasus positif.

$$\text{recal} = \frac{\text{TP}}{\text{TP+FN}} \qquad (2.6)$$

F1-score merupakan rata-rata harmonik dari precision dan recall yang dibobotkan. F1-score memiliki persentase akurasi yang lebih rendah dari accuracy karena menanamkan precision dan recall kedalam perhitungannya.

$$\text{f1score} = \frac{2 \text{ x precision x recal}}{\text{precision+recal}} \quad (2.7)$$

Accuracy digunakan untuk menghitung rasio dari prediksi yang benar, baik positif maupun negatif dengan keseluruhan data yang ada.

$$\text{accuracy} = \frac{(\text{TP+TN})}{\text{TP+FP+TN+FN}} \quad (2.8)$$

2.9 Bahasa Pemograman Python

Python adalah bahasa pemrograman interpretatif multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Python diklaim sebagai bahasa yang menggabungkan kapabilitas, kemampuan, dengan sintaksis kode yang sangat jelas, dan dilengkapi dengan fungsionalitas pustaka standar yang besar serta komprehensif (Syahrudin, 2018).

2.10 NumPy (Numerical Python)

NumPy atau Numerical Python adalah salah satu library teratas yang dilengkapi dengan sumber daya yang berguna untuk membantu para data scientist mengubah Python menjadi alat analisis dan pemodelan ilmiah yang kuat. Libary Open source terpopuler ini tersedia di bawah lisensi BSD. Ini adalah pustaka Python dasar untuk melakukan tugas dalam komputasi ilmiah. NumPy adalah bagian dari ekosistem berbasis Python yang lebih besar dari tool open source yang disebut SciPy (Adrianto, 2019).

2.11 Pandas

Pandas adalah library hebat lain yang dapat meningkatkan keterampilan Python untuk data science. Sama seperti NumPy, Pandas milik keluarga perangkat lunak open source SciPy dan tersedia di bawah lisensi perangkat lunak bebas BSD. Pandas menawarkan alat serbaguna dan kuat untuk struktur data dan melakukan analisis data yang luas. Library ini berfungsi dengan baik dengan data dunia nyata yang tidak lengkap, tidak terstruktur, dan tidak teratur - dan dilengkapi dengan tool untuk membentuk, menggabungkan, menganalisis, dan memvisualisasikan dataset (Adrianto, 2019).

2.12 Scikit-learn

Scikit-learn telah menjadi salah satu libraryopen source untuk machine learning paling populer di Python. Scikit-learn menyediakan algoritma untuk machine learning termasuk klasifikasi, regresi, dimensi reduksi, dan pengelompokan. Juga menyediakan modul untuk mengekstraksi fitur, memproses data, dan mengevaluasi model (Hackeling, 2017). Scikit-Learn atau SKleran Machine Learning in Python merupakan Alat sederhana dan efisien untuk penambangan data dan analisis data. Dapat diakses oleh semua orang, dan dapat digunakan kembali dalam berbagai konteks, dibangun di NumPy, SciPy, dan matplotlib, dan Sumber terbuka, dapat digunakan secara komersial - lisensi BSD (Ainun, 2021).

2.13 NLTK

NLTK (Naturel Language Toolkit) merupakan platform termuka dalam menciptakan program python untuk bekerja menggunakan data bahasa insan. Ini menyediakan anatar muka yang simpel di gunakan ke lebih dari 50 corpora dan lexical resources seperti WordNet, beserta

menggunakan rangkaian perpustakaan pemrosesan teks buat klasifikasi, tokenisasi, stemming, tagging, parsing dan penalaran sematik, perpustakaan NLP kekuatan industri dan forum diskusi yang aktif (Ainun, 2021).

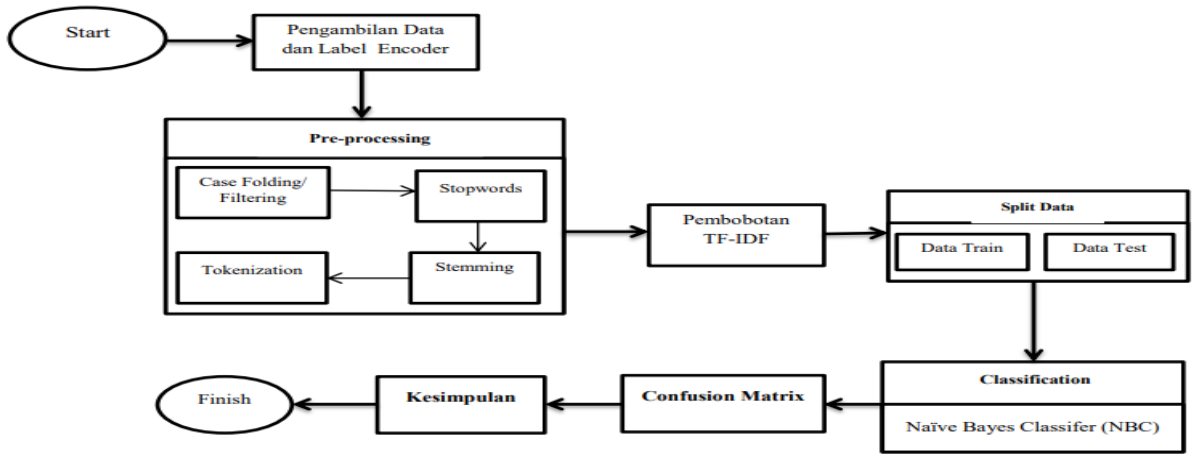
3. METODELOGI PENELITIAN

3.1 Teknik Pengumpulan data

Teknik Pengumpulan data yang dilakukan dengan crawler secara manual (copy- paste) dari Website resmi KOMINFO. Total berita yang diambil sebanyak 200 data yang muncul selama periode November 2021 – 2023. Kemudian dilakukan label encoder secara manual untuk memberikan label yang diklasifikasi kedalam hoax atau fakta. Dua value tersebut akan diubah setiap nilai dalam kolomnya menjadi angka yang berurutan, dimana untuk hoax akan dilabeli sebagai 0 dan fakta sebagai 1 yang nantinya akan menjadi acuan proses klasifikasi berita hoax untuk diujikan.

3.2 Diagram Alur Penelitian

Diagram alur penelitian ini bertujuan untuk menggambarkan proses dalam pengolahan data sesuai dengan rumusan masalah yang telah dipaparkan. Berikut tahapan diagram alur yang akan dilakukan dalam penelitian ini.



Gambar 2. Diagram Alur Penelitian

Langkah awal dalam penelitian ini adalah dengan pengambilan data pada Website Resmi KOMINFO dan pemberian Label Encover yaitu Hoax adalah 0 dan Fakta adalah 1. Langkah kedua melakukan Preprocessing adalah proses mempersiapkan teks agar dapat diolah lebih lanjut yang bertujuan untuk mengubah informasi dari setiap data ke dalam bentuk format yang lebih baku. Dalam tahap Prepocessing terdapat 4 tahap yaitu Case Folding/ Filtering, Stopwords, Steamming, dan Tokenization. Langkah ketigaMelakukan Pembobotan TF-IDF yaitu untuk mengubah teks menjadi angka supaya dapat diproses dalam Machine Learning. Langkah keempat adalah melakukan split data dengan membagi dataset menjadi 2 data data Train sebanyak 60% atau 140 data sedangkan data Test sebanyak 40% atau 80 Data.

Selanjutnya Classification dengan penerapan Metode Naïve Bayes Classifer (NBC) melakukan sebuah model dengan dasar data latih (Data Train) yang telah disiapkan, selanjnutnya dengan model tersebut akan dilakukan klasifikasi data yang baru. Tujuannya adalah agar sistem yang dibuat akan dapat melakukan pengklasifikasian data terhadap semua dataset dengan baik dan benar. Selanjutnya melakukan Confusion Matrix yang dapat memberikan informasi untuk merangkum kinerja dari klasifikasi yang sehubungan dengan beberapa test data yang divisualisasikan. Langkah Terakhir adalah Kesimpulan yang menjadi hasil dari Program yang sudah dijalankan dengan memasukkan data testing kedalam program akan menghasilkan klasifikasi terhadap data tersebut antara Hoax dan Fakta

4. HASIL DAN PEMBAHASAN

4.1 Penelitian Secara Umum

Dalam Penelitian ini, penulis akan mengimplementasikan metode data mining untuk Mendeteksi Berita Hoax Pada Dinas Kominikasi Informatika Dan PersandianPenulis akan menganalisis data menggunakan python dengan metode Naive Bayes Classifer (NBC).

4.2 Preprocessing

Preprocessing untuk membantu menghapusinformasi yang tidak relevan pada data dan dapat membantu mengurangi ukuran data mentahnya. Pada Kerja Praktek ini, tahapan preprocessing yang akan dilakukan diantaranya proses case Folding dengan merubah teks yang ada pada dataset menjadi huruf kecil (lower case), stopwords removal menyeleksi dan menghilangkan kata yang memiliki kemunculan tinggi pada dataset misalnya seperti kata penghubung, sapaan, kata yang tidak memiliki arti, dan sebagainya. stemming menghilangkan imbuhan yang ada pada tiap kata untuk mencari kata dasarnya dengan menggunakan library pada python yaitu Sastrawi yang cocok digunakan untuk menghilangkan imbuhan berbahasa Indonesia, dan terakhir yaitu tokenization dengan membagi kalimat yang semulanya lengkap menjadi baris huruf kecil. Berikut merupakan hasil dari tahap preprocessing seperti pada tabel 4.1.

No.	Tahap Preprocessing	Hasil
1.	Teks Berita	Beredar surat undangan seminar mengatasnamakan Direktor at Jenderal Pembelajaran dan Kemahasiswaan (Ditjen Belmaw a) Kementerian Pendidikan, Kebudayaan, Riset, dan Teknolog i (Kemendikbudristek). Surat tersebut bertema Seminar Buda ya Mutu Kepada Pimpinan Perguruan Tinggi dan ditujukan k epada pimpinan perguruan tinggi di seluruh Indonesia. Fakta nya, Lembaga Layanan Pendidikan Tinggi Wilayah IV Jawa B arat dan Banten melalui akun Instagram resminya @lldiktiwil ayah4, mengonfirmasi bahwa surat tersebut adalah hoaks. Pih aknya mengimbau masyarakat untuk waspada dan mengklari fikasi terlebih dahulu apabila mendapatkan informasi sejenis sebelum disebarkan.

2.	Case Folding /Filtering	beredar surat undangan seminar mengatasnamakan direktorat jenderal pembelajaran dan kemahasiswaan ditjen belmawa kementerian pendidikan kebudayaan riset dan teknologi kemendikbudristek surat tersebut bertema seminar budaya mutu kepada pimpinan perguruan tinggi dan ditujukan kepada pimpinan perguruan tinggi di seluruh indonesia faktanya lembaga layanan pendidikan tinggi wilayah iv jawa barat dan banten melalui akun instagram resminya lldiktiwilayah mengonfirmasi bahwa surat tersebut adalah hoaks pihaknya mengimbau masyarakat untuk waspada dan mengklarifikasi terlebih dahulu apabila mendapatkan informasi sejenis sebelum disebarkan
3.	Stopword	beredar surat undangan seminar mengatasnamakan direktorat jenderal pembelajaran kemahasiswaan ditjen belmawa kementerian pendidikan kebudayaan riset teknologi kemendikbudristek surat bertema seminar budaya mutu pimpinan perguruan pimpinan perguruan indonesia faktanya lembaga layanan pendidikan wilayah iv jawa barat banten akun instagram resminya lldiktiwilayah mengonfirmasi surat hoaks mengimbau masyarakat waspada mengklarifikasi informasi sejenis disebarkan
4.	Stemming	edar surat undang seminar mengatasnamakan direktorat jenderal ajar mahasiswa ditjen belmawa menteri didik budaya riset teknologi kemendikbudristek surat tema seminar budaya mutu pimpin guru pimpin guru indonesia fakta lembaga layan didik wilayah iv jawa barat banten akun instagram resmi lldiktiwilayah konfirmasi surat hoaks imbau masyarakat waspada klarifikasi informasi jenis sebar
5.	Tokenization	['edar', 'surat', 'undang', 'seminar', 'mengatasnamakan', 'direktorat', 'jenderal', 'ajar', 'mahasiswa', 'ditjen', 'belmawa', 'menteri', 'didik', 'budaya', 'riset', 'teknologi', 'kemendikbudristek', 'surat', 'tema', 'seminar', 'budaya', 'mutu', 'pimpin', 'guru', 'pimpin', 'guru', 'indonesia', 'fakta', 'lembaga', 'layan', 'didik', 'wilayah', 'iv', 'jawa', 'barat', 'banten', 'akun', 'instagram', 'resmi', 'lldiktiwilayah', 'konfirmasi', 'surat', 'hoaks', 'imbau', 'masyarakat', 'waspada', 'klarifikasi', 'informasi', 'jenis', 'sebar']

Tabel 1. Proses Preprocessing

4.3 Pembobotan Tf-Idf

Setelah melakukan preprocessing berupa Case folding, stopwords removal,stemming, tokenization, proses yang dilakukan selanjutnya yaitu TF-IDF (Term Frequency – Inverse

Document Frequency). TF-IDF dapat digunakan untuk mengetahui frekuensi dari istilah tertentu yang relatif terhadap sebuah kata dalam kumpulan dokumen dan melihat seberapa umum atau tidak umum sebuah kata yang ada diantara corpus (sekumpulan teks yang terstruktur). Pada proses TF-IDF ini menggunakan urutan token berupa unigram dalam implementasinya sehingga jumlah dari TF-IDF hanya satu kata saja.

Term	TF-IDF
edar	0.3099675219047144
media	0.3099675219047144
sosial	0.08061397731304756
postingan	0.08061397731304756
klaim	0.08061397731304756
bill	0.08061397731304756
gates	0.08061397731304756
masuk	0.05265263215509214
vaksin	0.05371751101877058
covid	0.065465135884297

Tabel 2 Hasil TF-IDF

Pada Tabel di atas Menampilkan hasil TF-IDF pada dokumen row dataframe ke –1 yang dapat dihitung dengan mengalikan dictionary dari TF-IDF secara value by value yang kemudian disimpan kedalam dataframe. Kemudian mengubah perhitungan series TF-IDF menjadi berbentuk Sparse Matrix.

4.4 Train Data Dan Test Data

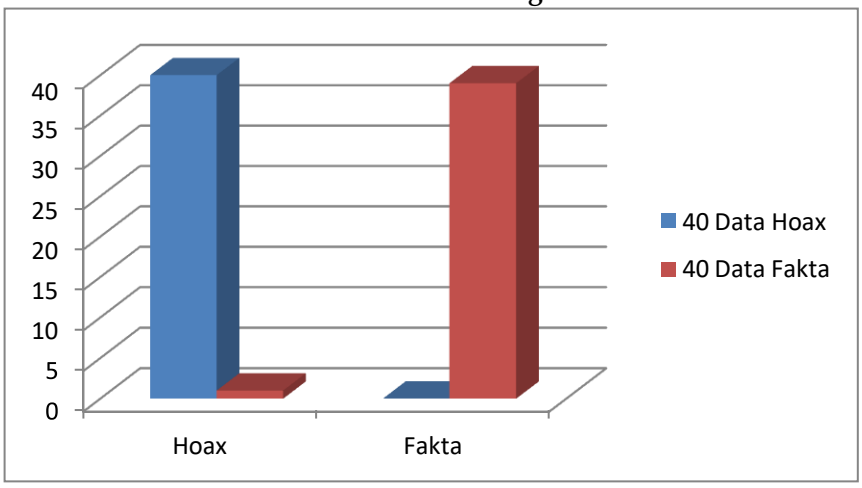
Train data atau data latih akan digunakan sebagai bahan dalam mencari model yang sesuai dengan cara melatih algoritma yang digunakan, sedangkan test data atau data uji nantinya akan digunakan untuk menguji dan mengetahui performa dari model algoritma Naïve Bayes Classifier yang diterapkan. Pada proses ini, jumlah data berita yang digunakan untuk diterapkan split data data sebanyak 200 dengan total berita 80 hoax dan berita fakta sebanyak 120 menggunakan perbandingan 60:40, 60% untuk train data dan 40% untuk test data.

4.5 Model Classifier

Dataset pada penelitian ini dibagi menjadi 2 set, train-set dan test-set untuk dimasukkan kedalam algoritma pengklasifikasian dengan perbandingan train-set 60% dan test-set 40% dari keseluruhan dataset. Kemudian data training digunakan sebagai acuan pada proses testing dengan algoritma Naïve Bayes Classifier untuk memastikan apakah data yang diuji telah tepat terhadap data testing. Data testing yang digunakan sebanyak 40 dataset hoax dan 40 dataset fakta dan di ambil secara acak. Berikut adalah hasil dari testing yang dilakukan secara manual.

Jumlah Data Testing	Hoax	Fakta
40 Data Hoax	40	0
40 Data Fakta	1	39

Tabel 3 Hasil Testing



Gambar 3. Diagram Batang Hasil Testing

Pada Tabel (3) dan Gambar (3) Menjelaskan bahwa sebanyak 80 dataset testing yang dilakukan sebanyak 79 data berhasil dan sesuai dengan class atau label yang digunakan dalam dataset yang digunakan. 40 dataset pada berita hoax yang diambil secara acak berhasil di testing dan tidak ada yang error. Sedangkan pada dataset fakta sebanyak 40 data yang ditesting sebanyak 39 data berhasil dan sesuai dengan dataset awal dan hanya 1 data yang error.

Dari modeling tersebut didapatkan hasil kinerja yang tampilkan dalam bentuk visualisasi confusion matrix dan classification report yang menunjukkan precision, recall, f1-score, support dan accuracy.

4.6 Hasil Evaluasi

Confusion Matrix memberikan informasi yang dapat digunakan untuk merangkum kinerja dari klasifikasi yang sehubungan dengan beberapa test data dengan memvisualisasikan. Hasil precision, recall, f1-score, support dan accuracy dengan menggunakan Algoritma Naïve Bayes Classifier adalah sebagai berikut :

```
test time:  0.769s
accuracy:   0.975
              precision    recall  f1-score   support

       Hoax       1.00      0.95      0.97        40
      Fakta       0.95      1.00      0.98        40

   accuracy                           0.97        80
  macro avg       0.98      0.97      0.97        80
weighted avg      0.98      0.97      0.97        80
```

Gambar 4. Hasil Akurasi Algoritma Naive Bayes Classifer

Setelah dilakukan penelitian dari tahap pengumpulan data hingga tahap model evaluation untuk mengklasifikasikan berita hoax dan fakta pada Diskominfo, dapat ditarik kesimpulan bahwa hasil performa rata-rata dari klasifikasi Naïve Bayes Classifier. Dengan nilai akurasi 0.975 atau 97% dengan nilai akurasi 97% menyatakan bahwa Algoritma Naïve Bayes Classifer efektif dalam mendeteksi berita hoax. Karena algoritma ini mengandalkan teorema Bayes untuk menghitung probabilitas berita hoax berdasarkan kemunculan kata- kata atau fitur-fitur lainnya dalam berita, dimana data yang digunakan pada penelitian sudah dilabeli terlebih dahulu dengan kelas hoax (0) dan fakta (1).

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil dari uraian dan pembahasan yang telah dilakukan, didapatkan kesimpulan yaitu mendeteksi hoax yang tersebar dapat dilakukan dengan melihat penggunaan kata pada berita juga lebih baku, narasumber yang dapat dipercaya serta biasanya bersifat informatif dan tanpa bias pada salah satu pihak saja, dan penelitian ini juga mampu untuk mengklasifikasikan berita hoax berbahasa Indonesia menggunakan algoritma Naïve Bayes Classifier serta melakukan evaluasi menggunakan TF-IDF. Dari hasil yang didapat dari algoritma Naïve Bayes Classifer adalah dengan nilai Akurasi 97% dengan Akurasi tersebut Algoritma Naïve Bayes Classifer efektif dalam mendeteksi berita Hoax pada Diskominfo.

5.2 Saran

Penelitian perlu kembangkannya lebih dimana memanfaatkan metode lain sehingga bisa dijadikan pembanding, menambah dataset supaya hasil yang diperoleh bisa lebih baik lagi dan mengembangkan program python tersebut dengan menggunakan website supaya lebih memudahkan dalam mengakses program pendeteksi berita hoax tersebut.

DAFTAR PUSTAKA

Adrianto, A. (2019). Analisis Cabutan Layanan Pada Pelanggan Indihome Menggunakan Algoritma Multiple Linear Regression Dengan Bahasa Pemrograman Python. Doctoral dissertation, Program Studi Informatika, Universitas Widyatama.

Ainun, N. (2021). analisis klasifikasi opini program studi menggunakan algoritma naïve bayes classifier pada universitas komputer indonesia. (Doctoral dissertation, Univeristas Komputer Indonesia).

Alfarisi, A. &. (2023). Penyebarluasan Berita Hoax Melalui Media Sosial (Studi Komparatif Pandangan Hukum Positif Indonesia dan Hukum Islam). Al-Amwal: Jurnal Ekonomi dan Perbankan Syariah.

ALVINA, D. (2023). Penerapan Kode Etik Jurnalistik Dalam Proses Produksi Berita Pada Radar Lampung (Doctoral Dissertation, Uin Raden Intan Lampung).

Baby, M. N. (2019). Customer classification and prediction based on data mining technique.

Cahyadi, R. (2020). Survei KIC: Hampir 60% Orang Indonesia Terpapar Hoax Saat Mengakses Internet. beritasatu.com.

Farach, D. &. (2019). Implementasi Metode Naïve Bayes Classifier dalam Analisis Sentimen Pada Opini Masyarakat Tentang RUU KUHP. Jurnal Advance in Social, Education and Humanities Research.

Fitria, U. E. (2018). erbandingan Kinerja Machine Learning Berbasis Algoritma Sepport Vectore Machine dan Naive Bayes. Skripsi Jurusan Statistika FMIPA UII.

ing, S. I. (2020). Is naive bayes a good classifier for document classification. International Journal of Software Engineering and Its Applications, 5, 37–46.

Juditha, C. (2018). Hoax Communication Interactivity in Social Media and Anticipation (Interaksi Komunikasi Hoax di Media Sosial serta Antisipasinya). Pekommas.

Kohli, S. (2019). "Understanding a Classification Report For Your Machine Learning .

Kusnandar, V. B. (2021). Pengguna Internet Indonesia Peringkat ke-3 Terbanyak di Asia. databoks.katadata.co.id.

Luqyana, W. A. (2018). Analisis Sentimen Cyrberbullyong pada Komentar Instangram dengan Metode Klasifikasi Support Vector Machine. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2, 4704-4713.

Mirzaalian, F. H. (2019). Social media analytics in hospitality and tourism and future trends. Journal of Hospitality and Tourism Technology,. https://doi.org/10.1108/ JHTT-08-2018-0078, 764–790.

Munjiah Nur Saadah, R. W. (2019). Sistem Temu Kembali Dokumen Teks dengan Pembobotan Tf-Idf Dan LCS. JUTI, 11, 17 – 20.

Syahrudin, A. N. (2018). Input dan output pada bahasa pemrograman python. Jurnal Dasar Pemograman Python STMIK, 20, 1-7.

Widiastuti, N. H. (2023). Komparasi Algoritma Klasifikasi Datamining Untuk Prediksi Minat Pencari Kerja. Jurnal Teknoinfo, 17, 219-227.