Journal Scientific of Mandalika (jsm) e-ISSN: 2745-5955, p-ISSN: 2809-0543, Vol. 6, No. 10, 2025
website: http://ojs.cahayamandalika.com/index.php/jomla
Accredited Sinta 5 based on SK. No. 177/E/KPT/2024

Latent Semantic Analysis (LSA) dengan Metode Support Vector Machine (SVM) dan Algoritma Naïve Bayes Pada Identifikasi Berita Palsu

Aliffia Putri Dito[1]*, Pulung Nurtantio Andono[2], M. Arief Soeleman[3].
[1,2,3]Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Indonesia
*Corresponding Author e-mail: aliffiadito@gmail.com

Abstract: *Fake news, or commonly known as hoaxes, are circulating widely in society. The spread of fake news can easily influence the public, especially through social media. The information disseminated through social media platforms is easily absorbed by the public. Social media users usually become content creators with a wide distribution of information, which allows for the presence of misinformation that cannot be ignored. The credibility of the information source is also crucial to avoid the risks of consuming fake news. According to statistical data published by Stanford University academics, 72.3 percent of fake news originates from social news outlets and online social media platforms. The identification of fake news is increasingly being utilized, but fact-checking is often challenging, time-consuming, and costly in many cases. This research was conducted using latent semantic analysis with support vector machine and naive Bayes algorithms in identifying fake news. The testing results yielded an accuracy rate of 82.28 percent using the support vector machine method and 81.39 percent with the naive Bayes algorithm.*
Keywords: *TF-IDF, LSA SVM, Naïve Bayes, Hoax*

Abstrack: Berita palsu atau disebut hoax banyak beredar di masyarakat. Penyebaran berita palsu dapat mudah menyerap masyarakat terlebih melalui media sosial. informasi yang tersebar melalui platform media sosial sangat mudah terserap bagi masyarakat. Para pengguna media sosial biasanya menjadi pembuat konten dengan jumlah penyebaran informasi yang cukup luas, dan memungkinkan adanya misinformasi yang tidak dapat diabaikan. Kredibilitas dari sumber informasi tersebut juga sangat penting untuk menghindari resiko mengkonsumsi berita palsu. Menurut data statistik yang diterbitkan oleh Stanford University academics, sebanyak 72,3 persen berita palsu berasal dari outlet berita sosial dan platform media sosial online. Identifikasi dalam berita palsu tersebut semakin meningkat penggunaannya namun pemeriksaan fakta dalam banyak kasus cukup sulit, memakan waktu dan memerlukan biaya yang besar. Penelitian ini dilakukan dengan menggunakan latent semantic analysis dengan metode support vector machine dan algoritma naïve bayes dalam identifikasi berita palsu. Hasil pengujian ini menghasilkan nilai akurasi sebesar 82,28% dengan metode support vector machine dan 81,39% pada algoritma naïve bayes.
Kata Kunci: TF-IDF, LSA SVM, Naïve Bayes, Hoax

Pendahuluan

Peningkatan penggunaan perangkat seluler dan akses internet yang mudah akhir-akhir membuat orang yang mengkonsumi informasi sama dengan penerimaan yang berbeda (Dunaway et al., 2018). Bermacam-macam media sosial seperti platform twitter dan platform instagram merupakan kanvas untuk informasi digital yang berkembang. Para pengguna media sosial biasanya menjadi pembuat konten dengan jumlah penyebaran informasi yang cukup luas, dan memungkinkan adanya misinformasi yang tidak dapat diabaikan. Kredibilitas dari sumber informasi tersebut juga sangat penting untuk menghindari resiko mengkonsumsi berita palsu (Nayoga et al., 2021).

Penelitian tentang identifikasi berita palsu ataupun deteksi berita palsu selalu berkembang. Panjaitan, dkk (Panjaitan & Santoso, 2021) melakukan penelitian dalam mendeteksi berita hoax berbahasa Indonesia tentang covid menyimpulkan dalam evaluasi model keseluruhan yang dilakukan, model paling baik adalah menggunakan algoritma random forest dengan fitur engineering. Penelitian ini mendapatkan nilai sebesar 96.05% pada akurasinya, 92.31% pada tingkat presisinya, 100% pada nilai sensitivitasnya, dan 96% pada f1-scorenya. Dan berdasarkan penelitian yang dilakukan oleh penulis tersebut, random forest dapat dikatakan yang paling efektif dalam klasifikasi berita hoaks berbahasa Indonesia. Faisal,

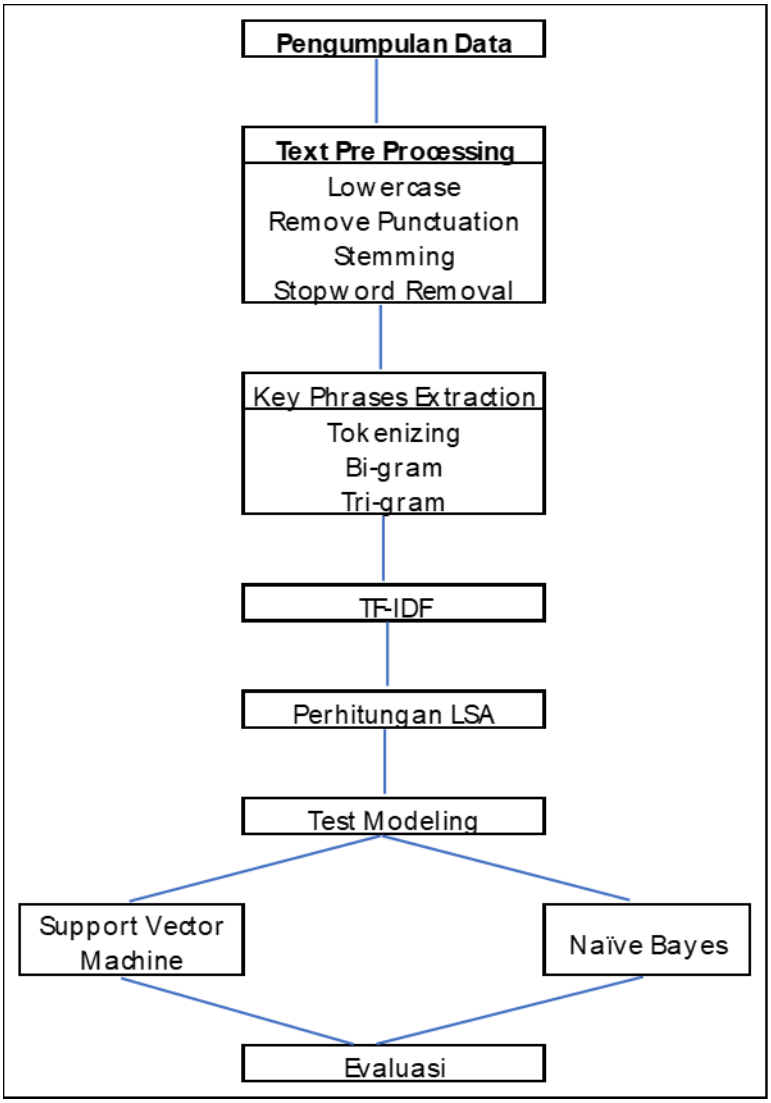
dkk (Rahutomo et al., 2019) mendeteksi kebenaran berita berita yang tersebar menggunakan metode naïve bayes, Sebanyak 600 berita telah diuji secara statis, dan hasil rata-rata akurasi pengujian tersebut adqalah 82,6%.

Penelitian lain oleh Kuai, dkk (Xu et al., 2019) melakukan deteksi berita palsu melalui media social dengan domain reputasi dan pemahaman konten. Penelitian ini mencirikan ratusan berita palsu populer dan berita kredibel popular dari berbagai perspektif termasuk domain dan reputasi penerbit berita, termasuk istilah penting dari setiap berita dan penyisipan kata. Penelitian ini menunjukkan bahwa berita palsu dan berita kredibel menunjukkan perbedaan substansi pada reputasi dan karakteristik domain penerbit berita tersebut. Di sisi lain, perbedaan pada topik dan penyematan kata menunjukkan perbedaan sedikit atau hampir tidak kentara pada berita palsu atau berita asli. Girgis, dkk (Girgis et al., 2018) pada penelitian algoritma deep learning dengan data besar untuk meningkatkan nilai pembelajaran dan mendapatkan hasil terbaik dengan menggunakan penyisipan kata untuk fitur ekstraksi atau isyarat yang membedakan hubungan antar kata dalam sintaksis dan semantic. Penelitian ini membahas implementasi model RNN (Vanilla, GRU) dan LSTMs untuk mendeteksi berita palsu online, dengan hasil menggunakan fitur GRU (Gated Recurrent Unit) adalah fitur terbaik karena dapat memecahkan masalah yang popular, serta CNN (Convolutional Neural Network) adalah model terbaik karna kecepatan dan hasil kinerja terbaiknya untuk windows.

Penelitian yang dilakukan oleh Alfiyah, dkk (Jamaludin, 2022) mengembangkan sistem dalam penelitian deteksi informasi palsu di platform media sosial menggunakan fitur ekspans Glove) dengan menggunakan tiga klasifikasi SVM, NB dan RNN. Hasil penelitian dapat meningkatkan akurasi sebesar 91,92% dengan fitur ekspansi yang digunakan dalam penelitian tersebut.

Pada penelitian ini akan dilakukan identifikasi berita palsu sebagai tujuan utama. Masalah yang melatar belakangi penelitian ini adalah berita palsu yang semakin meningkat di masyarakat dalam beberapa tahun belakangan. Sayangnya, belum ada definisi yang disepakati tentang istilah berita palsu. Untuk memandu arah penelitian identifikasi berita palsu dengan lebih baik, diperlukan klasifikasi yang tepat. Media sosial merupakan sumber yang kuat dalam penyebaran berita berita bohong. Ada beberapa pola yang muncul yang dapat dimanfaatkan untuk identifikasi berita palsu di media sosial. Tinjauan tentang metode pendeteksian berita palsu pada penelitian-penelitian sebelumnya dari berbagai skenario media sosial dapat memberikan pemahaman dasar tentang metode pendeteksian berita palsu yang canggih. Deteksi berita palsu di media sosial masih dalam tahap awal pengembangan, dan masih banyak isu menantang yang perlu diselidiki lebih lanjut. Penting untuk membahas arah penelitian potensial yang dapat meningkatkan kemampuan deteksi dan mitigasi berita palsu.

Metode Penelitian



Gambar 1 Usulan Metode

Penelitian ini menggunakan dataset public yang berasal dari Kaggle, diproses menggunakan Text pre-processing meliputi lowercase, remove punctual, stemming, stopword removal dan pembobotan kata menggunakan TF-IDF, setelah itu dilakukan key phrases extraction meliputi tokenizing, bi-gram dan tri-gram, selanjutnya akan dilakukan pembobotan kara menggunakan TF-IDF, dan akan dilakukan perhitungan LSA. Setelah itu akan di test modeling menggunakan SVM dan Naïve Bayes untuk mendapatkan hasil.

1. Pengumpulan Data

Penelitian ini menggunakan data yang berasal dari Kaggle (https://www.kaggle.com/datasets/muhammadghazimuharam/indonesiafalsenews). Data set yang digunakan diperoleh dari data Kaggle, yang berisikan data latih dan data uji dari berita palsu berbahasa Indonesia tahun 2018 sampai 2020, jumlah record data latih sebanyak 4.231 dan record data test sebanyak 470 dengan 2 atribut. Data yang digunakan ini berupa data set berita palsu dan berita valid. Adapun penjabaran atribut dijelaskan pada Tabel 1. sebagai berikut:

Tabel 1 Tabel Atribut

No	Atribut	Keterangan
1	Judul	Merupakan judul dari berita yang tampil pada data
2	Narasi	Merupakan narasi pertama dari berita tersebut

2. Text Pre processing dan Key Phrase Extraction

Setelah melakukan pengumpulan data dari Kaggle. Selanjutkan akan di lakukan text pre processing pada dataset. Text pre-processing yang digunakan dalam penelitian ini adalah lowercase, remove punctual, stemming, stopword removal dan pembobotan kata menggunakan TF-IDF.

a. Lowercase

Dataset yang sudah dimiliki akan melalui tahap lowercase terlebih dahulu. Lowercase adalah proses perubahan seluruh kata dan huruf pada dataset yang awalnya huruf kapital menjadi huruf kecil atau non-kapital.

b. Remove Punctuation

Setelah data dilakukan lowercase. Data akan di Remove Punctuation, yang merupakan proses penghapusan karakter dalam sebuah kata yang harusnya tidak diperlukan, seperti emotikon, angka, tanda baca (.); (,); (!); (?), alamat pada sebuah url (https);(www.com); hastag (#), dan nama pengguna (@namaanda) yang terdapat pada dataset.

c. Stopword Removal

Data yang sudah di remove punctual kan di Stopward removal, dengan mengurangi atau menghilangkan kata yang tidak sesuai dan yang tidak memiliki arti khusus seperti kata ganti, preposisi dan konjungsi. Proses ini dilakukan dengan tujuan untuk dilakukan proses penghilangan kata-kata umum yang biasanya muncul namun tidak memiliki makna atau akan mengambil kata-kata yang dianggap penting.

d. Tokenizing

Data yang sudah dilakukan stopword removal akan dilakukan proses Tokenizing. Dalam proses tokenisasi, teks input dipotong menjadi kata, istilah, symbol, tanda baca atau elemen lain yang memiliki makna yang disebut (Vijayarani & Janani, 2016). Tanda baca seperti titik(.), koma(,), tandaseru(!), tanda tanya(?) dan sejenisnya dianggap tidak diperlukan dalam proses tokenisasi dan akan dihilangkan.

e. Stemming

Stemming adalah proses mengubah kata yang diekstraksi menjadi bentuk dasarnya dengan cara menghilangkan imbuhan kata.(Panjaitan & Santoso, 2021). Penerapan stemming dalam setiap bahasa bervariasi tergantung pada struktur morfologi bahasa tersebut. Tujuan utama dari proses ini adalah untuk memahami maksa suatu kata meskipun telah memiliki bentuk yang berbeda. Data yang sudah dilakukan preses tokenizing hasilnya akan dilakukan proses steeming.

f. Bi-Gram

Model bahasa yang didasarkan pada penentuan probabilitas berdasarkan jumlah urutan dua kata.

g. Tri-Gram

Model bahasa yang didasarkan pada penentuan probabilitas berdasarkan jumlah urutan tiga kata.

3. Latent Semantic Analysis

Pengacuan pustaka dalam naskah ini menggunakan gaya penulisan APA (American Psychological Association), di mana sitasi dalam teks ditulis dengan mencantumkan nama belakang penulis dan tahun terbit, seperti (Andono, 2021) atau (Dito, Andono, & Soeleman, 2020). Setiap sumber yang disitasi dalam naskah harus dicantumkan dalam daftar pustaka, dan sebaliknya, setiap entri dalam daftar pustaka harus memiliki sitasi dalam teks. Daftar pustaka

disusun secara alfabetis berdasarkan nama belakang penulis, bukan berdasarkan urutan kemunculan dalam teks. Disarankan untuk menggunakan perangkat manajemen referensi seperti Mendeley atau Zotero untuk membantu dalam pengelolaan kutipan dan pembuatan daftar pustaka sesuai gaya APA.

4. Klasifikasi menggunakan SVM

Setelah proses pemilihan fitur menggunakan TF-IDF dan LSA selesai, dilakukan tahap klasifikasi untuk membangun model dengan menggunakan SVM Non Linear. Klasifikasi dengan SVM dilakukan dengan mengkonfigurasi penggunakaan kernel dari hasil pembelajaran data sebelumnya. Tahap ini dilakukan dua kali eksperimen. Eksperimen pertama yaitu SVM tanpa seleksi fitur, dan selanjutnya eksperimen dengan seleksi fitur. Klasifikasi dilakukan dengan menggunakan data training dengan split data menjadi 3, dan dilakukan klasifikasi SVM Non Linear dengan one-gram, bi-gram, tri-gram.

5. Klasifikasi menggunakan Naïve Bayes

Setelah proses pemilihan fitur menggunakan TF-IDF dan LSA selesai, dilakukan tahap klasifikasi untuk membangun model dengan menggunakan Gaussian Naïve Bayes. Tahap ini dilakukan dua kali eksperimen. Eksperimen pertama yaitu Gaussian Naïve Bayes tanpa seleksi fitur, dan selanjutnya eksperimen dengan seleksi fitur. Klasifikasi dilakukan dengan menggunakan data training dengan split data menjadi 3, dan dilakukan klasifikasi Gaussian Naïve Bayes dengan one-gram, bi-gram, tri-gram.

6. Pengujian dan Evaluasi Hasil Penelitian

Tujuan dari pengujian ini adalah untuk mempresentasikan hasil dari data yang telah diuji. Pada tahap pengujian ini dilakukan perhitungan akurasi. Pengujian eksperimen yang dilakukan berupa pengukuran tingkat akurasi klasifikasi berita palsu. Identifikasi berita hoax dengan menggunakan bantuan machine learning merupakan hal yang sudah umum untuk dilakukan.

Pada tahap penelitian ini, dilakukan pengujian model dengan menggunakan Teknik cross validation dengan 10 folds, dimana tiap proses ini membagi data ke dalam 10 bagian secara acak. Nilai yang dihasilkan dari pengujian ini berupa klasifikasi dari data pengujian yang dipilih sebelumnya. Untuk menghitung performa klasifikasi dengan SVM dan Naïve Bayes ini pengukuran yang dipakai adalah tingkat akurasi dengan menggunakan metode one-gram, bi-gram dan tri-gram. Nilai tersebut didapatkan dari variable pada table Confusion Matrix. Untuk mengetahui tingkat keberhasilan penelitian, nilai akurasi dari tahapan tanpa seleksi fitur akan dibandingkan dengan tahap pemilihan fitur (TF-IDF+LSA)

Evaluasi dan validasi hasil yang didapatkan adalah table pengukuran Confusion Matrix. Perhitungan akurasi pada penelitian ini dilakukan dengan menggunakan rumus dibawah ini.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

Dimana:

TP = Prediksi positif yang positif	FP = Prediksi negatif yang positif
FN = Prediksi positif yang negatif	TN = Prediksi negatif yang negatif

Hasil dan Pembahasan

Berdasarkan dari metodologi penelitian, proses diatas menghasilkan beberapa hasil.

1. Pengumpulan Data

Pada penelitian dalam identifikasi berita palsu kali ini menggunakan data set yang digunakan diperoleh dari data Kaggle, yang berisikan data latih dan data uji dari berita palsu berbahasa Indonesia tahun 2018 sampai 2020, jumlah record data latih sebanyak 4.231 dan record data test sebanyak 470 dengan 2 atribut.

2 Text Pre processing

Tahapan Text Pre Processing dilakukan agar data yang akan digunakan menjadi data yang bersih untuk dapat di proses. Pre-processing yang dilakukan antara lain lowercase, remove punctual, stopword removal, tokenizing, stemming, setelah dilakukan stemming data akan dilakukan proses ekstraksi key phare dengan menggunakan bi-gram dan tri-gram. Proses text preprocessing menggunakan aplikasi Phyton.

3 Pembobotan Kata menggunakan TF-IDF

Setelah tahapan text pre processing selesai, akan dilakukan pembobotan text menggunakan TF-IDF. Perhitungan TF-IDF yang akan dituliskan dalam table dibawah merupakan perhitungan manual TF-IDF dari Dataset Judul. Term Frequency (TF) yang akan digunakan untuk menghitung frekuensi sebuah kata muncul pada dataset judul dan IDF- Invers Document Frequency untuk pembobotan pada kata tertentu yang banyak terkandung dalam sebuah dokumen. Implementasi TF-IDF menggunakan Library Scikit-learn (Pedregosa et al., 2011).

Pada tahapan ini dilakukan pembobotan fitur-fitur data yang sudah diproses tokenizing, bigram dan tri-gramnya. Dengan melakukan perhitungan fungsi TF-IDF Term Weighting dari persamaan dibawah sehingga menghasilkan nilai TF-IDF tiap fitur.

$$W(i,j) = TF(i,j).IDF \quad (11)$$

$$IDF = \log\left(\frac{N}{dfj}\right) \quad (12)$$

Dimana:

a) $TF(i,j)$ adalah total kemunculan kata ke-i dalam dokumen ke-j
b) $N\ adalah\ jumlah\ total\ dokumen$
c) dfj adalah banyak dokumen yang mengandung kata ke-i

Untuk mendapatkan nilai invers document frequency (IDF) masing-masing dokumen pada setiap term, akan dilakukan perhitungan dari persamaan (12) dengan mengambil sample term ‘tilang’.

$$IDF = \log\left(\frac{N}{dfj}\right) = \log\left(\frac{2}{5}\right) =$$

Untuk mendapatkan bobot (W) masing-masing dokumen pada setiap term, akan dilakukan perhitungan dari persamaan (11) dengan mengambil sample term ‘tilang’.

$$W(i,j) = TF(i,j).IDF = \times 0{,}397 = 0{,}795$$

Sebagai contoh perhitungan TF-IDF dibawah ini, merupakan dataset judul yang sudah dilakukan tahapan text pre-processing.

Tabel 2. Perhitungan Manual TF-IDF

Term	TF					DF	tfdt= df/ND	IDF= log(df/ND)	TF-IDF
	D1	D2	D3	D4	D5		5		
pakai	1	0				1	0,2	-0,69897	-0,13979
masker	1	1				2	0,4	-0,39794	-0,15918
sebab	1					1	0,2	-0,69897	-0,13979
sakit	1					1	0,2	-0,69897	-0,13979

legionnaries	1					1	0,2	-0,69897	-0,13979
instruksi		1				1	0,2	-0,69897	-0,13979
gubernur		1				1	0,2	-0,69897	-0,13979
tentang		1				1	0,2	-0,69897	-0,13979
tilang		2				2	0,4	-0,39794	-0,15918
bagi		1				1	0,2	-0,69897	-0,13979
tidak		1				1	0,2	-0,69897	-0,13979
muka		1				1	0,2	-0,69897	-0,13979
umum		1				1	0,2	-0,69897	-0,13979
guna		1				1	0,2	-0,69897	-0,13979
apps		1				1	0,2	-0,69897	-0,13979
pikobar		1				1	0,2	-0,69897	-0,13979
foto			1		2	3	0,6	-0,22185	-0,13311
jim			1			1	0,2	-0,69897	-0,13979
rohn			1			1	0,2	-0,69897	-0,13979
jokowi			1	1		2	0,4	-0,39794	-0,15918
adalah			1			1	0,2	-0,69897	-0,13979
presiden			1			1	0,2	-0,69897	-0,13979
baik			1			1	0,2	-0,69897	-0,13979
sejarah			1			1	0,2	-0,69897	-0,13979
bangsa			1			1	0,2	-0,69897	-0,13979
indonesia			1			1	0,2	-0,69897	-0,13979
ini				1	1	2	0,4	-0,39794	-0,15918
bukan				1		1	0,2	-0,69897	-0,13979
politik				1		1	0,2	-0,69897	-0,13979
nyata				1		1	0,2	-0,69897	-0,13979
berhasil				1		1	0,2	-0,69897	-0,13979
pulang				1		1	0,2	-0,69897	-0,13979
triliun				1		1	0,2	-0,69897	-0,13979

uang				1		1	0,2	-0,69897	-0,13979
negara				1		1	0,2	-0,69897	-0,13979
swiss				1		1	0,2	-0,69897	-0,13979
kadrun					1	1	0,2	-0,69897	-0,13979
lihat					1	1	0,2	-0,69897	-0,13979
panas					1	1	0,2	-0,69897	-0,13979
dingin					1	1	0,2	-0,69897	-0,13979

4 Latent Semantic Analysis – LSA

Setelah dataset yang sudah melalui proses pembobotan kata menggunakan TF-IDF, akan digunakan algoritma kedua untuk mencari dan menemukan informasi berdasarkan keseluruhan makna dokumen, bukan hanya makna antar individu kata menggunakan Latent Semantic Analysis. Implementasi LSA menggunakan Library Scikit-learn (Pedregosa et al., 2011).

Sebagai contoh sample dari sebuah data $A = [2,0,0]; [2,1,0]; [0,-2,0]$ akan dilakukan proses Latent Semantic Analysis - LSA, dengan tahapan:

1) Membentuk Inputan Matriks A

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & -2 & 0 \end{bmatrix}$$

2) Melakukan dekomposisi matriks A diatas menjadi menjadi tiga komponen matriks yang lebih sederhana.
   a. Menghitung nilai matriks $N = A^T A$

$$\begin{bmatrix} 2 & 2 & 0 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 2 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & -2 & 0 \end{bmatrix} = \begin{bmatrix} 8 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

   b. Mencari nilai eigenvalue $Det(N - \lambda I)$

$$Det\left(\begin{bmatrix} 8 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \lambda I\right) = \det\left(\begin{bmatrix} 8 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \left(\lambda \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right)\right)$$

$$= \det\left(\begin{bmatrix} 8-\lambda & 2 & 0 \\ 2 & 5-\lambda & 0 \\ 0 & 0 & -\lambda \end{bmatrix}\right)$$

$$\det(N - \lambda I) = (-1)^2(8-\lambda)\begin{vmatrix} 5-\lambda & 0 \\ 0 & -\lambda \end{vmatrix} + (-1)^3(2)\begin{vmatrix} 2 & 0 \\ 0 & -\lambda \end{vmatrix} + (-1)^4(0)\begin{vmatrix} 0 & 0 \\ 0 & 0 \end{vmatrix}$$

$$= (8-\lambda)(-5\lambda + \lambda^2) + 4\lambda$$

$$= \lambda^3 - 5\lambda^2 - 36\lambda = \lambda(\lambda - 4)(\lambda - 9)$$

$$\lambda_1 = 9\ ;\ \lambda_2 = 4;\ \lambda_3 = 0$$

Setelah mendapatkan nilai eigenvalue nya berupa $\lambda_1 = 9\ ;\ \lambda_2 = 4;\ \lambda_3 = 0$

   a. Menghitung nilai $\sigma_i = \sqrt{\lambda_i}$

$$\sigma_1 = \sqrt{9} = 3;\ \sigma_2 = \sqrt{4} = 2; \sigma_3 = 0$$

   b. Membentuk Matriks Σ

$$\sum = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

c. Membentuk Matriks $U = AV$

5. Perhitungan Naïve Bayes Gaussian

Setelah dataset sudah dilakukan proses latent semantic analysis – lsa akan dilakukan proses perhitungan naïve bayes gaussian dengan contoh sebagai berikut :

Tabel 3. Contoh Data Perhitungan Naive Bayes

No	Komponen 1	Komponen 2	Komponen3	Label
1	3,50E-11	-3,19E-11	-1,15E-08	TRUE
2	-1,66E-11	5,11E-11	2,50E-08	TRUE
3	4,20E+01	-7,38E+01	3,88E+04	TRUE
4	3,02E-11	6,06E-12	1,31E-08	TRUE
5	-4,83E-11	1,41E-11	-6,59E-08	FALSE
6	4,57E-12	-2,91E-12	3,66E-08	TRUE
7	-1,45E-11	-6,04E-12	-1,02E-08	FALSE
8	-1,62E-11	3,99E-12	-2,43E-08	FALSE
9	-8,53E-12	-1,02E-11	-1,85E-08	TRUE
10	2,11E-11	-6,24E-11	-2,54E-09	TRUE

Tabel 4. Contoh Dataset

TEST 1	6,02E-02	4,43E+00	-2,75E+00
TEST 2	4,03E-11	-3,03E-10	-2,40E-07
TEST 3	3,96E-11	-5,10E-11	2,24E-08
TEST 4	-2,16E-11	2,86E-11	1,64E-08
TEST 5	2,89E-10	4,72E-10	-3,37E-07

Menghitung Rata-rata untuk Komponen 1 Berlabel True, dengan rumus sebagai berikut

$$\mu_{1,T} = \frac{\sum_{i=1}^{n} xi}{n} = \frac{3{,}50 + (-1{,}66) + 4{,}2 + 3{,}02 + 4{,}56 + (-8{,}52) + 2{,}11}{7} = 1{,}03$$

Hasil perhitungan nilai rata-rata untuk Komponen 2 berlabel true sebesar $\mu_{2,T}$ =1,05, dan hasil perhitungan nilai rata-rata untuk Komponen 3 berlabel true sebesar $\mu_{3,T}$ =5,55.

Menghitung Standart Deviasi untuk Komponen 1 Berlabel True, dengan rumus sebagai berikut

$$\delta^2{}_{1,T} = \frac{\sum_{i=1}^{n} (xi - \mu)^2}{n-1}$$

$$= \frac{(3{,}5-1{,}03)^2+(-1{,}66-1{,}03)^2+(4{,}2-1{,}03)^2+(3{,}02-1{,}03)^2+(4{,}56-1{,}03)^2+(-8{,}52-1{,}03)^2+(2{,}11-1{,}03)^2}{6} = 1{,}59$$

Hasil perhitungan nilai standar deviasi untuk Komponen 2 berlabel true sebesar $\delta^2{}_{2,T} = 2{,}79$, dan hasil perhitungan nilai standar deviasi untuk Komponen 3 berlabel true sebesar $\delta^2{}_{3,T} = 1{,}47$.

Menghitung Rata-rata untuk Komponen 1 Berlabel False, dengan rumus sebagai berikut

$$\mu_{1,F} = \frac{\sum_{i=1}^{n} xi}{n} = \frac{((-4{,}83) + (-1{,}45) + (1{,}62))}{3} = -1{,}55$$

Hasil perhitungan nilai rata-rata untuk Komponen 2 berlabel false sebesar $\mu_{2,F} = 4{,}01$,

dan hasil perhitungan nilai rata-rata untuk Komponen 3 berlabel false sebesar $\mu_{3,F}$ = -3,35.

Menghitung Standart Deviasi untuk Komponen 1 Berlabel False, dengan rumus sebagai berikut

$$\delta^2{}_{1,F} = \frac{\sum_{i=1}^{n}(xi-\mu)^2}{n-1} = \frac{((-4{,}83+1{,}55)^2+(-1{,}45+1{,}55)^2+(1{,}62-1{,}55)^2}{2} = 5{,}38$$

Hasil perhitungan nilai standar deviasi untuk Komponen 2 berlabel false sebesar $\delta^2{}_{2,F} = 2{,}79$, dan hasil perhitungan nilai standar deviasi untuk Komponen 3 berlabel false sebesar $\delta^2{}_{3,F} = 1{,}47$.

Selanjutnya akan dihitung nilai fungsi likelihood menggunakan data test 1 pada table 4.8 dengan rumus sebagai berikut

$$P(x_1|TRUE) = \frac{1}{\sqrt{(2\pi\sigma^2)}}e^{\frac{(x-\mu)^2}{2\delta^2}} = \frac{1}{\sqrt{(2(3{,}14)(1{,}59))}}e^{\frac{(6{,}02-1{,}03)^2}{2(1{,}59)}} = (-0{,}965)$$

$$P(x_1|FALSE) = \frac{1}{\sqrt{(2\pi\sigma^2)}}e^{\frac{(x-\mu)^2}{2\delta^2}} = \frac{1}{\sqrt{(2(3{,}14)(5{,}38))}}e^{\frac{(6{,}02+1{,}55)^2}{2.(5{,}38)}} = 35{,}35$$

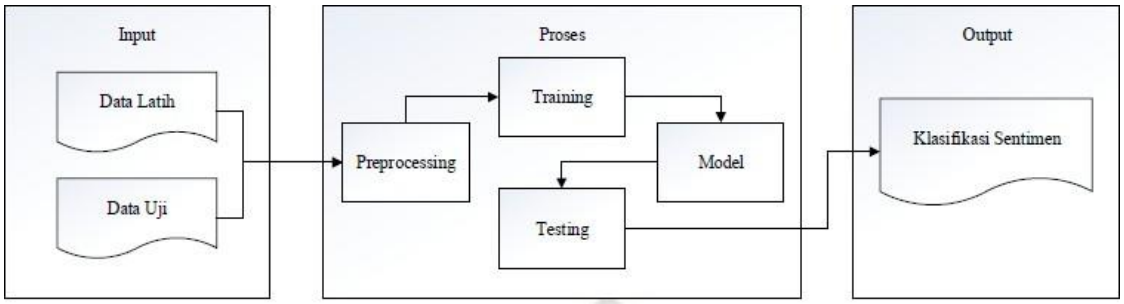
Setelah dicari nilai dari fungsi likelihoodnya, akan dicari probabilitas dari dataset table 4.8.

6. Perhitungan Support Vector Machine

Proses SVM akan melibatkan training menggunakan data latih untuk menghasilkan model pembelajaran dengan menggunakan k-fold cross validation, serta menggunakan parameter C dan Gamma. Model terbaik merupakan model yang memiliki akurasi nilai tertinggi dari pasangan C dan Gamma terbaik yang diperoleh dari Teknik grid searching dan dengan menggunakan 10-fold cross validation. Pada proses testing menggunakan data uji untuk mengklasifikasikan data uji tersebut.

7. Pembelajaran dan Model

Pembelajaran pada SVM menggunakan data latih untuk mendapatkan model klasifikasi pada SVM. Tahap pembelajaran ini melibatkan fungsi kernel sebagai fungsi transformasi. Kernel yang digunakan adalah kernel RBF. Kernel RBF membutuhkan pasangan parameter C dan Gamma. Untuk mendapatkan nilai parameter C dan Gamma terbaik dilakukan dengan metode grid search dan 10-fold cross validation. Metode grid search bertujuan untuk membuat grid parameter dengan cara menentukan nilai untuk parameter C dan Gamma secara manual. Menetukan nilai parameter C dan menetukan nlai parameter Gamma, dan melakukan pencarian grid untuk C dan Gamma dengan cara menerapkan 10-fold cross validation . 10-fold cross validation adalah pembagian data latih menjadi 10 segmen sama banyak. Kemudian akan dilakukan 10 kali proses training dan testing dengan perbandingan 9/10 segmen sebagai data latih dan 1/10 segmen sebagai data uji. Selanjutnya menghitung rataan akurasi untuk keseluruhan fold. Kemudian melakukan pemilihan parameter C dan Gamma terbaik berdasarkan akurasi paling tinggi. Berikut merupakan tahapan klasifikasi SVM secara umum dapat dilihat pada gambar dibawah.



Gambar 2.Tahapan Klasifikasi SVM

1) Input

Data yang akan diinputkan adalah keseluruhan data komentar yang telah dikumpulkan sebanyak 1000 dan telah dilakukan proses pelabelan secara manual dengan bantuan ahli bahasa. Data akan terbagi menjadi beberapa skenario yang akan menghasilkan validasi akurasi tertinggi.

2) Proses SVM

Merupakan tahapan mulai dari preprocessing, pembobotan sampai pembentukan pemodelan. Dalam proses ini juga sudah dilakukan validasi akurasi menggunakan k-fold dan confusion matrix.

3) Output

Data yang akan dihasilkan adalah berupa data yang telah diprediksi oleh pemodelan svm dengan kelas positif dan negatif.

8. Evaluasi Hasil

Penelitian ini menggunakan dataset berita palsu sebanyak 4231 record data. Pengujian ini dilakukan terhadap 3 dataset, yang pertama dataset one-gram, yang kedua dataset bi-gram, yang ketiga dataset tri-gram. Hasil perancangan sistem ini akan dimulai dengan tahapan text preprocessing antara lain lowercase, remove punctual, stopword removal, tokenizing dan stemming dengan library sastrawi. Setelah dilakukan text preprocessing akan dilanjutkan dengan tahapan pemberian bobot dengan menggunakan TF-IDF menggunakan library scikit-learn (Pedregosa et al., 2011) dan menerapkan implementasi latent semantic analysis menggunakan library scikit-learn (Pedregosa et al., 2011).

Berikut adalah nilai akurasi dari pengujian SVM Non-Linear yang sudah dilakukan proses TF-IDF dan LSA:

Tabel 5. Nilai accuracy SVM

SVM Non Linear	Dataset 1	Dataset 2	Dataset 3
One-Gram	0.82440945	0.82047244	0.82362205
Bi-Gram	0.82204724	0.81968504	0.81732283
Tri-Gram	0.81968504	0.81889764	0.81889764

Rata-rata nilai accuracy One-gram $= \frac{0.82440945+0.82047244+0.82362205}{3} = 0{,}822834647$

Rata-rata nilai accuracy Bi-gram=0,819685037

Rata-rata nilai accuracy Tri-gram = 0,819160107

Tabel 6. Nilai accuracy Naive Bayes

Naïve Bayes G	Dataset 1	Dataset 2	Dataset 3
One-Gram	0.81811024	0.80944882	0.81417323
Bi-Gram	0.38188976	0.79370079	0.25748031
Tri-Gram	0.18346457	0.18031496	0.18267717

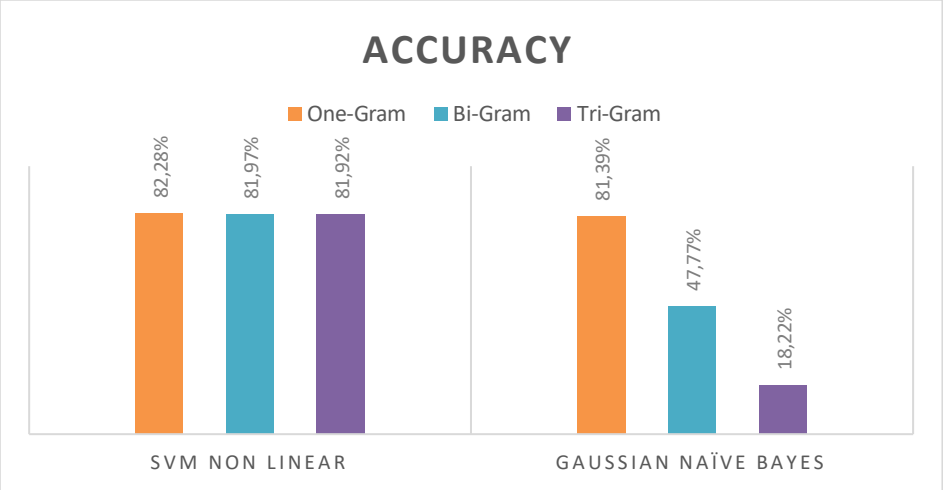
Rata-rata nilai accuracy One-gram sebesar 0,813910763. Rata-rata nilai accuracy Bi-gram sebesar 0,477690287. Rata-rata nilai accuracy Tri-gram sebesar 0,182152233

Tabel 7. Hasil Perbandingan Accuracy menggunakan dataset yang sudah dilakukan proses TF-IDF dan LSA

	One-Gram	Bi-Gram	Tri-Gram
SVM Non Linear	82,28%	81,97%	81,92%
Gaussian Naïve Bayes	81,39%	47,77%	18,22%

Berdasarkan table 7 ditunjukkan nilai accuracy dengan SVM Non Linear dan Gaussian Naïve Bayes. Nilai akurasi terendah sebesar 18,22% dihasilkan dengan algoritma gaussian naïve bayes dengan tri-gram dan hasil akurasi tertinggi sebesar 82,28% dengan metode support vector machine dengan one-gram.

Hasil yang didapatkan dari pengujian menggunakan SVM Non-Linear dan Gaussian Naïve Bayes pada table 7 kemudian disusun ke dalam diagram seperti gambar berikut:



Gambar 3. Hasil Accuracy

Berdasarkan hasil accuracy pada gambar 3. diatas, metode gaussian naïve bayes dengan dataset one-gram mendapat nilai 81,39%, sedangkan pada dataset bi-gram dan tri-gram mendapatkan hasil akurasi yang rendah yaitu 47,77% dan 18,22%, hal ini dapat disebabkan karena jumlah dataset yang kecil dan terbatas. Keterbatasan ini dapat menyebabkan kurangnya variasi dan representasi yang cukup dari jenis-jenis bi-gram dan tri-gram yang ada dalam teks. Dataset yang terbatas ini mengakibatkan kesulitan menangkap pola dan hubungan yang kompleks antar kata-kata dalam teks yang lebih besar.

Sedangkan untuk metode SVM non linear pada dataset one-gram mendapatkan nilai 82,28%, dan dengan metode SVM non linear pada dataset bi-gram mendapatkan nilai 81,97%, dan pada dataset tri-gram mendapatkan nilai 81,92%.

Kesimpulan dan Saran

Berdasarkan penelitian yang sudah dilakukan pada identifikasi berita palsu menggunakan latent semantic analysis dengan metode support vector machine dan naïve bayes dapat disimpulkan sebagai berikut.

1. Dengan diterapkannya beberapa proses pada text preprocessing seperti lowercase, remove punctual, stopword removal, tokenizing dan stemming dan juga pembobotan

kata menggunakan TF-IDF, mampu mengurangi dimensi jumlah kata pada data dan membantu kecepatan pada proses klasifikasi.

2. Metode klasifikasi menggunakan Support Vector Machine Non Linear dengan menerapkan TF-IDF+LSA pada kategori dataset one-gram, bi-gram dan tri-gram menghasilkan akurasi sebesar 82,28%, 81,97%, 81,92% secara berurutan. Hal ini disebabkan oleh dataset yang kecil memiliki jumlah data yang terbatas. Keterbatasan ini dapat menyebabkan kurangnya variasi dan representasi yang cukup dari berbagai jenis bi-gram dan trigram yang ada dalam teks. Akibatnya, model yang dilatih pada dataset kecil mungkin tidak dapat menangkap pola dan hubungan yang kompleks antara kata-kata dalam teks yang lebih besar. ada dataset kecil, terdapat risiko overfitting, yaitu ketika model terlalu spesifik dan terlalu cocok dengan data pelatihan yang terbatas. Overfitting dapat menghasilkan hasil akurasi yang tinggi pada data pelatihan, tetapi performa yang buruk saat digunakan pada data baru. Model yang overfit mungkin mempelajari hubungan yang khusus untuk dataset kecil tersebut, yang tidak dapat umumkan untuk dataset yang lebih besar atau berbeda. Dataset yang kecil mungkin tidak mewakili secara merata seluruh variasi dan distribusi kata-kata dalam teks yang lebih besar. Hal ini dapat menyebabkan beberapa bi-gram dan trigram memiliki frekuensi yang sangat rendah atau bahkan tidak ada dalam dataset kecil tersebut. Sebagai hasilnya, model mungkin tidak dapat belajar dengan baik tentang kombinasi kata-kata yang jarang muncul atau memiliki pola yang tidak biasa. Dalam text mining, konteks sangat penting dalam memahami arti kata-kata dan hubungan antara mereka. Dataset bi-gram dan trigram yang kecil mungkin tidak memberikan konteks yang cukup untuk kata-kata yang muncul dalam teks.
3. Metode klasifikasi menggunakan Gaussian Naïve Bayes dengan menerapkan TF-IDF+LSA pada dataset menghasilkan akurasi sebesar 81,39% pada kategori dataset one-gram, sebesar 47,77% pada kategori dataset bi-gram dan 18,22% pada kategori data tri-gram.

Berdasarkan hasil pengujian yang telah dillakukan dan hasil kesimpulan yang sudah dijabarkan diatas, perlu adanya saran dan usul dalam pengembangan penelitian ini, antara lain:

1. Pengujian dapat dilakukan dengan metode lain agar dapat melihat perbandingan metode dan mencari metode yang paling efektif digunakan
2. Untuk penelitian selanjutnya disarankan untuk mengumpulkan dataset yang lebih besar dan lebih bervariasi, serta menggunakan teknik word embeddings yang telah dilatih sebelumnya untuk mengenali pola dan hubungan kata yang lebih kompleks.

Rekomendasi

Berdasarkan hasil penelitian yang dilakukan dalam artikel ini, terdapat beberapa rekomendasi penting untuk penelitian selanjutnya yang dapat meningkatkan efektivitas dan akurasi dalam identifikasi berita palsu. Pertama, disarankan agar penelitian selanjutnya menggunakan dataset yang lebih besar dan beragam, karena keterbatasan jumlah data pada penelitian ini terbukti mempengaruhi kinerja model, khususnya pada penggunaan bi-gram dan tri-gram yang memiliki akurasi rendah. Dataset yang lebih besar akan memungkinkan model untuk menangkap lebih banyak variasi dan konteks kata yang dibutuhkan dalam mendeteksi berita palsu secara lebih akurat. Kedua, penggunaan teknik representasi kata yang lebih

canggih seperti word embedding (contohnya: Word2Vec, GloVe, atau BERT) perlu dipertimbangkan agar model dapat memahami makna semantik dan konteks antar kata secara lebih mendalam. Selain itu, pengujian lebih lanjut terhadap algoritma pembelajaran mesin lainnya, seperti Random Forest, XGBoost, atau pendekatan deep learning seperti LSTM dan CNN, juga layak dieksplorasi untuk mengetahui apakah ada metode lain yang dapat menghasilkan kinerja lebih baik dibandingkan SVM dan Naïve Bayes. Akhirnya, penting untuk melakukan analisis performa tidak hanya berdasarkan akurasi, tetapi juga melalui metrik evaluasi lain seperti precision, recall, dan F1-score, agar gambaran performa model menjadi lebih komprehensif. Dengan demikian, penelitian di masa depan diharapkan dapat memberikan solusi yang lebih andal dan aplikatif dalam menangani penyebaran berita palsu di era digital saat ini.

Referensi

Dunaway, J., Searles, K., Sui, M., & Paul, N. (2018). News attention in a mobile era. *Journal of Computer-Mediated Communication*, *23*(2), 107–124.

Girgis, S., Amer, E., & Gadallah, M. (2018). Deep learning algorithms for detecting fake news in online text. *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 93–97.

Jamaludin, A. R. (2022). *Deteksi berita hoax di media sosial twitter dengan ekspansi fitur menggunakan glove*.

Nayoga, B. P., Adipradana, R., Suryadi, R., & Suhartono, D. (2021). Hoax analyzer for Indonesian news using deep learning models. *Procedia Computer Science*, *179*, 704–712.

Panjaitan, A. T. B., & Santoso, I. (2021). Deteksi Hoaks Pada Berita Berbahasa Indonesia Seputar COVID-19. *Jurnal FORMAT (Teknik Informatika)*, *10*(1), 76.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Rahutomo, F., Pratiwi, I. Y. R., & Ramadhani, D. M. (2019). Eksperimen naïve bayes pada deteksi berita hoax berbahasa Indonesia. *Jurnal Penelitian Komunikasi Dan Opini Publik*, *23*(1).

Vijayarani, S., & Janani, R. (2016). Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, *3*(1), 37–47.

Xu, K., Wang, F., Wang, H., & Yang, B. (2019). Detecting fake news over online social media via dXu, K., Wang, F., Wang, H., & Yang, B. (2019). Detecting fake news over online social media via domain reputations and content understanding. Tsinghua Science and Technology, 25(1), 20–27.omain reputations. *Tsinghua Science and Technology*, *25*(1), 20–27.