Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893
p-ISSN: 2089-3787

Pembangunan Fitur dalam Identifikasi Cerdas Hoaks dengan *Naïve Bayes* dan Klasifikasi *Decision Tree*

Muhammad Umar Shalih[1], Teja Endra Eng Tju[2*]
Teknologi Informasi, Universitas Budi Luhur, DKI Jakarta, Indonesia
*e-mail *Corresponding Author:* teja.endraengtju@budiluhur.ac.id

Abstract

Identifying hoaxes poses significant complexity and challenges due to issues such as the diverse nature of hoaxes, rapid narrative changes, swift dissemination, sophisticated technological usage, verification difficulties, and scalability challenges. Recognizing the societal impact of hoaxes, the development of features for intelligent hoax identification research becomes imperative. The methodology adopted from CRISP-DM and SKKNI No. 299 of 2020, customized to research needs, encompasses five stages: data understanding, data preparation, modeling, evaluation, and deployment. Data from Mafindo comprises 9,756 instances divided into 7,804 training data and 1,952 test data. Six features source, capital, keyword, sentiment, fact-check, and classification are utilized as supervisory labels. Sentiment and fact-check features are constructed using the Multinomial Naïve Bayes method and modeled using the Decision Tree technique on the dataset. Modeling variations include dataset quantities of 2,000, 4,000, 6,000, and 8,000, along with addressing imbalance dataset issues. Utilizing the Confusion Matrix technique, modeling results indicate an accuracy of 93.5% and an F1 score of 0.935. It's observed that the imbalanced dataset minimally affects accuracy and F1 score but contributes to model stability concerning the quantity of data with specific labels.

Keywords: *Classification and Regression Trees; SMOTE; Confusion Matrix; Fact Check; Mafindo*

Abstrak

Identifikasi hoaks cukup kompleks dan menantang dengan permasalahan seperti keanekaragaman hoaks, perubahan narasi yang cepat, kecepatan penyebaran yang luas, penggunaan teknologi canggih, kesulitan verifikasi, dan tantangan skala, yang dihadapi. Sebagai kepedulian dampak hoaks pada masyarakat, penelitain pembangunan fitur dalam identifikasi cerdas hoaks perlu dilakukan. Metodologi diadopsi dari CRISP-DM dan SKKNI No. 299 tahun 2020 yang disesuaikan dengan kebutuhan penelitian sehingga menjadi lima tahapan yaitu *data understanding, data preparation, modeling, evaluation,* dan *deployement*. Data diperoleh dari Mafindo dan digunakan sebanyak 9.756 data yang dibagi menjadi 7.804 data latih dan 1.952 data uji. Terdapat enam fitur yaitu sumber, kapital, *keyword*, sentimen, *factcheck*, dan klasifikasi sebagai label supervisi. Dua fitur sentimen dan *factcheck* dibangun dengan metode *Multinomial Naïve Bayes*, selanjutnya dilakukan pemodelan pada *dataset* dengan metode *Decision Tree*. Pemodelan dilakukan pula dengan variasi kuantitas *dataset* sebanyak 2.000, 4.000, 6.000, 8000, juga dengan perbandingan masalah imbalance *dataset*. Hasil pemodelan dengan teknik *Confusion Matrix* diperoleh akurasi 93,5% dan skor F1 0,935 dan diperoleh bahwa *imbalance dataset* tidak terlalu berpengaruh pada hasil akurasi dan skor F1 namun memberikan kestabilan model dalam hal kuantitas besarnya data dengan label tertentu.

Kata kunci: *Classification and Regression Trees; SMOTE; Confusion Matrix; Fact Check; Mafindo*

1. Pendahuluan

Identifikasi hoaks merupakan hal yang kompleks dan menantang. Perkembangan teknologi informasi dan media sosial melalui internet telah memberikan akses yang luas bagi individu untuk menyebarkan informasi secara cepat dan mudah [1]. Namun, dampak negatif dari kemudahan ini adalah penyebaran hoaks (berita bohong) yang dapat menyesatkan dan mempengaruhi opini publik. Dari perspektif hukum pidana, pelaku penyebaran hoaks juga dapat

di pidana [2], sebagai pertanggungjawaban hukum individu yang sengaja menyebarkan atau menciptakan berita palsu dengan niat menyesatkan dan merugikan orang lain.

Di tengah arus informasi yang berkembang pesat, gap antara kemampuan identifikasi hoaks yang efektif dan kecepatan penyebarannya semakin melebar. Beberapa permasalahan yang sering dihadapi dalam identifikasi hoaks adalah: keanekaragaman hoaks dalam berbagai bentuk dan konten melalui berbagai *platform* dan saluran komunikasi; pola perubahan yang cepat dalam narasi, strategi, dan teknik manipulasi informasi; kecepatan penyebaran dapat menjangkau ribuan atau bahkan jutaan orang dalam waktu singkat; penggunaan teknologi seperti kecerdasan buatan, pemrosesan bahasa alami, atau teknik manipulasi gambar dan video yang canggih; kesulitan verifikasi terutama dalam konteks berita yang cepat berubah atau kontroversial, mencari sumber terpercaya, memverifikasi fakta, dan mendapatkan konteks yang tepat membutuhkan waktu dan upaya yang signifikan; tantangan skala dengan pendekatan skala besar, termasuk penggunaan teknologi dan kecerdasan buatan, untuk mengatasi tantangan ini. Hal ini menciptakan kebutuhan akan pendekatan baru yang mampu menangani hoaks dengan lebih efisien dan akurat.

Penelitian ini menggarap isu serius terkait hoaks di Indonesia dengan kolaborasi bersama Mafindo. Menghadapi kompleksitas tantangan ini, diperlukan pengembangan pendekatan baru yang mampu menangani berbagai permasalahan identifikasi hoaks yang ada. Melalui pemanfaatan referensi dan metodologi yang kuat, penelitian ini bertujuan untuk mengusulkan solusi yang lebih efektif dalam mengidentifikasi hoaks. Pendekatan yang diusulkan berfokus pada pengembangan fitur-fitur identifikasi hoaks yang didasarkan pada analisis mendalam dari data berita *online* berasal dari berbagai sumber. Fitur-fitur yang dibangun berdasarkan pada kredibilitas sumber berita [3], [4], penggunaan huruf besar [5], kata kunci (kontribusi dari Mafindo), sentimen negatif [6]–[8], *hoax factcheck* [9]–[11].

Tujuan penelitian untuk mengisi kesenjangan dalam identifikasi hoaks dengan pendekatan yang lebih canggih dan terstruktur, yang diharapkan akan memberikan kontribusi dalam upaya memerangi penyebaran hoaks di Indonesia. Melalui analisis yang mendalam, penelitian ini juga diharapkan dapat memberikan pandangan yang lebih komprehensif terhadap permasalahan hoaks dan solusi yang diusulkan.

2. Tinjauan Pustaka

Tinjauan pustaka dilakukan dengan mencari, membaca, dan mempelajari artikel jurnal yang telah dilakukan oleh para peneliti. Artikel yang ditinjau difokuskan pada teknik klasifikasi hoaks menggunakan *machine learning* [12], [13], dengan ringkasan hasil tinjauan ditunjukkan pada Tabel 1. Tampak bahwa penelitian terkait hoaks kebanyakan didominasi metode kalsifikasi *Naïve Bayes* dengan hasil akurasi 72,00% sampai dengan 98,50% [8], [14]–[26], sedangkan metode lainnya adalah SVM (*Support Vector Machine*) dengan akurasi 90,70% [27] hingga 95,60% [28], KNN (*K-Nearest Neigbour*) dengan akurasi 75,40% [29] hingga 75,89% [30], *Random Forest* dengan akurasi 76,47% [31], dan *Decision Tree* dengan akurasi 72,91% [32]. Data berita ada yang berasal dari satu atau dua sumber tertentu atau dari berbagai sumber yang dituliskan sebagai Artikel Berita. Jumlah data bervariatif dari 30 saja sampai 50.610. Topik berita ada yang mengacu pada satu topik tertentu, beberapa topik, atau tanpa memperhatikan topik.

Tabel 1. Ringkasan Tinjauan Pustaka

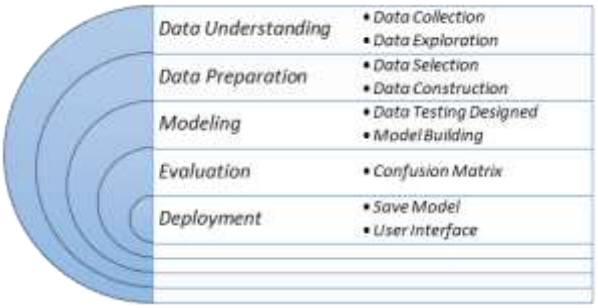
Penulis	Sumber (Jumlah Data)	Metode	Akurasi	Topik
[32]	Twitter (50.610)	*Decision Tree*	72,91%	6 topik
[30]	Jabar Saber Hoaks, Jala Hoaks (599)	KNN	75,89%	Covid-19
[29]	Artikel Berita (740)	KNN	75,40%	-
[14]	Kaggle (11.000)	*Naïve Bayes*	98,50%	-
[15]	detik.com, turnbackhoax.id (300)	*Naïve Bayes*	93,00%	-
[16]	turnbackhoax.id (110)	*Naïve Bayes*	91,82%	-
[17]	Artikel Berita (220)	*Naïve Bayes*	91,36%	22 topik
[8]	Artikel Berita (30)	*Naïve Bayes*	91,00%	-
[18]	Twitter (720)	*Naïve Bayes*	88,80%	Covid-19
[19]	liputan6.com, turnbackhoax.id (300)	*Naïve Bayes*	86,30%	Covid-19
[20]	turnbackhoax.id (300)	*Naïve Bayes*	85,28%	Covid-19
[21]	turnbackhoax.id (150)	*Naïve Bayes*	85,19%	-
[22]	detik.com, turnbackhoax.id (1.849)	*Naïve Bayes*	85,09%	6 topik

Penulis	Sumber (Jumlah Data)	Metode	Akurasi	Topik
[23]	Artikel Berita (600)	*Naïve Bayes*	82,60%	-
[24]	data.mendeley.com (600)	*Naïve Bayes*	82,00%	-
[25]	kumparan.com (1.000)	*Naïve Bayes*	81,00%	-
[26]	Artikel Berita (250)	*Naïve Bayes*	72,00%	10 topik
[33]	Twitter (12.000)	*Neural Network*	97,30%	Covid-19
[34]	Twitter (50.646)	*Neural Network*	78,76%	6 topik
[31]	Artikel Berita (251)	*Random Forest*	76,47%	-
[28]	Artikel Berita (287)	SVM	95,60%	Kesehatan
[27]	Kaggle (20.008)	SVM	90,70%	-

Berbeda dengan penelitian sebelumnya, dalam penelitian ini digunakan enam fitur yaitu sumber berita, huruf besar, kata kunci, sentimen negatif, *hoax factcheck*, dan klasifikasi berita (sebagai label). Fitur klasifikasi berita diperoleh secara langsung dari data berita yang diambil dari Mafindo (Masyarakat Anti Fitnah Indonesia) [35], sedangkan fitur lainnya hasil rekayasa dari isi berita tersebut. Dua fitur sentimen dan *factcheck* dikonstruksi dengan teknik *Multinomial Naïve Bayes*, sedangkan metode prediksi untuk klasifikasi digunakan *Decision Tree* [36] dengan algoritma CART (*Classification and Regression Trees*) [37] karena lebih fleksibel pada independensi fitur dan tipe data, serta banyaknya data yang diolah.

3. Metodologi

Metodologi diadopsi dari CRISP-DM (*CRoss Industry Standard Process for Data Mining*) [38] dan SKKNI (Standar Kompetensi Kerja Nasional Indonesia) No. 299 tahun 2020 [39] yang disesuaikan dengan kebutuhan penelitian sehingga menjadi lima tahapan yaitu *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployement*, seperti ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Penelitian

3.1. *Data Understanding*

Tahapan penelitian dimulai dengan pemahaman data mulai dari pengambilan data hingga eksplorasi data.

1) *Data Collection*

Data berita dalam format JSON (*JavaScript Object Notation*) [40], diambil dengan API (*Application Programming Interface*) yang disediakan dan diberikan akses oleh Mafindo, selanjutnya data disimpan dalam database *MySQL* dan ditampilkan dengan *phpMyAdmin*.

2) *Data Exploration*

Setiap berita terdiri dari atribut *Title* (judul berita)*, Content* (isi berita)*, Source* (sumber berita)*,* dan *Classification* (klasifikasi berita yang berisi label *Valid* atau *Hoax*), seperti sampel berita yang disajikan pada Gambar 2.

Title	Content	Source	Classification
[SALAH] "CIA Bongkar jati diri Presiden Indonesia ...	"Di nilai Indonesia condong ke Blok China Komunis...	facebook.com	Hoax
[SALAH] Hidayat Nur Wahid Akui PKS Tak Menganut As...	HIDAYAT NUR WAHID AKUI PKS TAK MENGANUT ASAS PANCA...	facebook.com	Hoax
[SALAH] "sakitnya apa di diagnosa nya covid"	Beredar sebuah foto dan akun Nanda Faturachman (f...	facebook.com	Hoax
Sebaran Kasus COVID-19 Di Kabupaten Temanggung	Jika sudah terjangkit covid-19 didaerah mana saja...	whatsapp.com	Valid
FAQ Wabah Virus Corona Apa Saja yang Perlu Kita T...	Adakah corona virus bisa menebak melalui permukaan...	whatsapp.com	Valid
[SALAH] "Apa iyaaaaa bisa hamil????"	Beredar postingan sejumlah foto pria tengah melaku...	facebook.com	Hoax
46 Tenaga Medis RS Kanadi Positif Corona, Lanjut ...	Dokter dan perawat rumah sakit karyadi positif cov...	whatsapp.com	Valid
[SALAH] "bupati Luwu Utara terbaring di RS karena ...	Beredar foto bupati Luwu Utara terbaring di RS kar...	facebook.com	Hoax
[SALAH] Alhamdulillah Akhirnya Aceh Bisa Berangkat...	"Alhamdulillah Akhirnya Aceh Bisa Berangkat Haji ...	youtube.com	Hoax
[SALAH] Wapres Pemerintah Gak Sengaja Memakai Danz...	Akun Facebook Putra Inka membagikan gambar judul s...	facebook.com	Hoax
[SALAH] "dokter gigi di Surabaya stres telanjang d...	Akun facebook Samuel Moeladi mengunggah sebuah g...	facebook.com	Hoax
[SALAH] "Ibu Kota Pindah, Anies Bakal Jual Gedung ...	Beredar artikel berjudul "Ibu Kota Pindah, Anies B...	facebook.com	Hoax
[SALAH] "3 hari kedepan Arus angin dari Utara ke a...	Arus angin dari utara dan utara ke arah selatan y...	whatsapp.com	Hoax
[SALAH] Minggu 21 Juni 2020 Baim Wong Rilis Hadiah...	Beredar postingan Facebook dengan gambar poster ar...	facebook.com	Hoax
[SALAH] "Amien Rais: Kalau Jokowi Sampai Diturunka...	Akun Afrizal (fb.com/100035229611489) mengunggah s...	facebook.com	Hoax
[SALAH] Artikel "#MakzulkanJKWBubarkanPDIP Trendin...	#MakzulkanJKWBubarkanPDIP Trending, RUU HIP Ditudi...	facebook.com	Hoax
[SALAH] Tulisan "Mana teriakan "Saya Pancasila mu?"...	Mana teriakan "Saya Pancasila mu ?" Oleh : Dr Abd...	facebook.com	Hoax
Tiga Orang Dekat Menpan RB Tjahjo Kumolo Positif C...	Tjahjo Kumolo positif covid-19 breaking news	whatsapp.com	Valid
[SALAH] Kapal Tenggelam di Perairan Makassar, 13 O...	Innalillahi wa innailaihi Rojuun. Turut berduka c...	whatsapp.com	Hoax

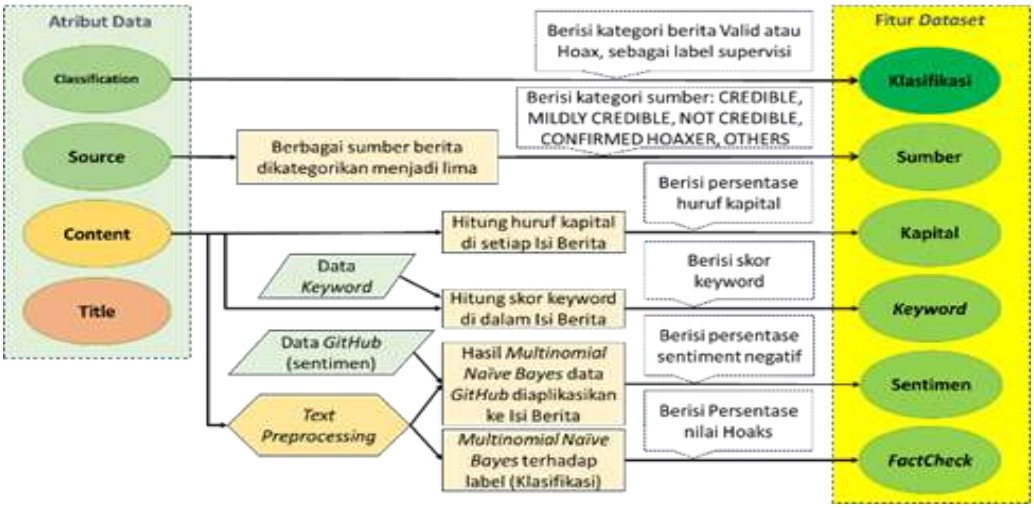
Gambar 2. Sampel Data dari Mafindo

3.2. *Data Preparation*

Persiapan data dengan pemilihan dan rekayasa dari atribut data menjadi fitur-fitur yang diperlukan sehingga didapat *dataset* untuk pemodelan.

1) *Data Selection*

Data dianalisa dan ditentukan atribut-atribut yang akan digunakan sebagai fitur pada *dataset* pemodelan. Fitur klasifikasi diambil secara langsung dari atribut *classification*. Fitur klasifikasi ini merupakan label yang digunakan sebagai supervisi, sedangkan fitur lainnya diolah dari atribut *source* dan *content*. Pemilihan dan transformasi fitur ini bisa dilihat pada Gambar 3.



Gambar 3. Pemilihan dan Konstruksi Fitur

2) *Data Construction*

Atribut *source* yang berisi berbagai data media online dikategorikan menjadi lima (*Credible, Mildly Credible, Not Credible, Confirmed Hoaxer, Others*). Dua kategori pertama diadopsi berdasarkan tingkatan status perusahaan pers [41] yaitu berturut-turut "Terverifikasi Administrasi dan Faktual" dan "Terverifikasi Administrasi", sedangkan *Not Credible* adalah sumber yang tidak terdaftar di dewan pers. Comfirmed Hoaxer adalah sumber berita yang masuk dalam daftar hoaks Mafindo. Untuk *Others* dikategorikan untuk berita-berita yang tidak diketahui sumbernya.

Selain fitur Sumber, empat fitur dikonstruksi dari atribut isi berita dan diberi nama Kapital, *Keyword*, Sentimen, dan *Factcheck*. Fitur kapital diperoleh dari persentase banyaknya huruf besar/kapital dalam setiap isi berita. Fitur *keyword* berupa total skor dari kata kunci hoaks (milik Mafindo) yang muncul dalam berita. Fitur sentimen dihasilkan dengan *Multinomial Naive Bayes* menggunakan data kata sentimen positif [42] dan negatif [43] dari *repository* di *GitHub* dan diaplikasikan pada setiap isi berita sehingga diperoleh dan digunakan persentase sentimen negatif. Demikian juga untuk fitur *FactCheck* dihasilkan dengan *Multinomial Naïve Bayes* pada hasil *text preprocessing* seluruh isi berita kemudian terhadap label dilakukan perhitungan probabilitas hoaks setiap berita sehingga diperoleh dan digunakan persentase hoaks. Konstruksi fitur juga disajikan pada Gambar 2.

3.3. *Modeling*

Pemodelan dilakukan dengan terlebih dahulu *dataset* dipisahkan menjadi data latih dan data uji. Semua proses perhitungan menggunakan bahasa pemograman *Python* [44], [45] dengan *Visual Studio Code* (*VSCode*) sebagai editor.

1) *Data Testing Designed*

Total data yang diperoleh sebanyak 11.934 , namun ada 2.178 data yang belum ada label (klasifikasi) karena belum dilakukan pemeriksaan fakta oleh tim Mafindo, sehingga *dataset* yang dipakai terdiri dari total 9.756 data yang terdiri dari 8.940 hoaks dan 816 valid. *Dataset* dibagi menjadi 80% data latih dan 20% data uji, sedangkan 2.178 data tanpa label tidak digunakan. Pelatihan juga dilakukan dengan jumlah data random yang bervariasi untuk mengetahui pengaruh besarnya data dengan pemodelan *Decision Tree Classifier* CART [37], dengan distribusi data disajikan pada Tabel 2. Untuk membuktikan pengaruh *imbalance dataset*, pelatihan dilakukan pula dengan dan tanpa teknik SMOTE [46].

Tabel 2. Percobaan dengan Variasi Jumlah Data

Total Data Pemodelan	***Dataset***	
	Data Latih	**Data Uji**
9.756	7.804	1.952
8.000	6.400	1.600
6.000	4.800	1.200
4.000	3.200	800
2.000	1.600	400

2) *Model Building*

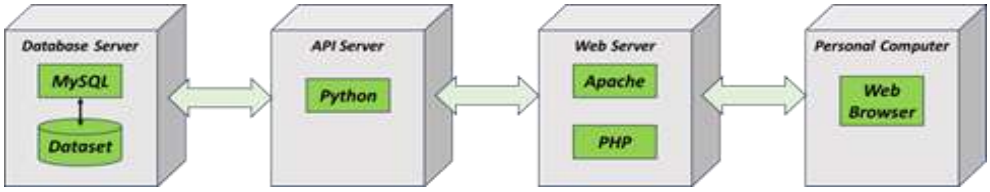
Pemodelan dilakukan dengan metode *Decision Tree Classification* [36] yang mengimplementasikan algoritma CART.

3.4. *Evaluation*

Hasil pemodelan dievaluasi dengan *confusion matrix* berikut *accuracy* dan F1 *Score* sebagai metrik pengukurannya [47] pada beberapa *dataset* seperti di Tabel 2.

3.5. *Deployment*

Proses dan hasil pemodelan diimplementasikan berbasis *web* dengan arsitektur seperti pada Gambar 4.



Gambar 4. Deployment Environment

1) *Save Model*

Hasil pemodelan, digunakan *pickle library* untuk diubah menjadi *binary file*, disimpan ke dalam *storage API Server*, kemudian dengan *Python API* diumpankan ke *web server*.

2) *User Interface*

Aplikasi pengguna dibuat dengan *HTML* dan *CSS* yang diakses melalui *web browser*.

4. Hasil dan Pembahasan

Setelah dilakukan tahapan dalam metodologi penelitian, didapatkan hasil yang secara komprehensif dijelaskan ke dalam empat tahap berikut ini.

4.1. *Data Preparation*

Data untuk fitur Sumber diperoleh dengan algoritma pada Gambar 5.

```
Calculate Score Sumber(string sumber)

source_rankings = get_source_rankings()
sumber_input = lower_case(sumber)

for each source_ranking in source_rankings do
   list_source = source_ranking['list_source']
   the score of the source = source_ranking['ranking']

   for each source in list_source do
      if source is found in sumber_input then
         return the score of the source
      end if

return source is not found
```

Gambar 5. Algoritma Data Fitur Sumber

Untuk mendapatkan data fitur Kapital dari data atribut *Content* digunakan algoritma yang ditunjukkan pada Gambar 6.

```
Calculate Score Capital(string teks)

count_capital := 0
count_lower := 0
total_lower_capital := 0

for each character c in teks do
   if c is uppercase then
      increment count_capital by 1
   else if c is lowercase then
      increment count_lower by 1
   end if

set total_lower_capital as the sum of count_capital and count_lower

if total_lower_capital is 0 then
   set percentage_capital as 0
else
   set percentage_capital as count_capital divided by total_lower_capital
end if

return percentage_capital
```

Gambar 6. Algoritma Data Fitur Kapital

Selanjutnya fitur *Keyword*, data diperoleh dari atribut *Content* dan daftar kata kunci hoaks dari Mafindo menggunakan algoritma seperti pada Gambar 7.

```
Calculate Score Keyword (list KEYWORDS_SCORES, string teks)

score_teks = 0

for each keyword and score from KEYWORDS_SCORES items do
   occurrence = find occurrence of keyword in teks
   score_teks += (occurrence * score)

return score_teks
```

Gambar 7. Algoritma Data Fitur *Keyword*

Sebelum mendapatkan data untuk fitur Sentimen dan *FactCheck*, dilakukan *text preprocessing* data pada atribut *Content* dengan algoritma di Gambar 8.

```
PreProcess Text(string teks)

CaseFoldedText = caseFolding(teks)
SymbolRemovedText = symbolRemoval(teks)
TokenizedWords = WordTokenization(teks)
StopWordRemovedText = stopWordRemoval(teks)
ClearSingleCharacterText = removeSingleCharacterWords(teks)
CleanText = removeWhitespace(teks)

Return CleanText
```

Gambar 8. Algoritma *Text Preprocessing*

Dengan algoritma pada Gambar 9, dilakukan *Multinomial Naïve Bayes* dengan data kata sentimen [42], [43] dari *GitHub,* hasil pemodelannya digunakan algoritma pada Gambar 10 untuk menghitung probabilitas negatif pada hasil *text preprocessing* data atribut *Content* untuk mengisi data fitur Sentimen.

```
Train Data Sentimen ()

negative_keywords = open file of negative keywords
positive_keywords = open file of positive keywords

keywords = negative_keywords combined with positive_keywords
labels = add 0 for count of negative_keywords times
labels += add 1 for count of positive_keywords times

mnb = create MultinomialNaiveBayes object
mnb.fit(keywords, labels)

store mnb in a binary file using pickle
```

Gambar 9. Algoritma Pemodelan Data Sentimen *GitHub*

```
Predict Sentimen With Probability (string teks)

trained_model = load trained model from file (FILE_SENTIMEN_MODEL)

text to predict =  preprocessing(teks)

new_text_features =  Count Vectorize (new_text)
probabilities = trained_model.Predict With Probabilities(new_text_features)
predictions  = trained_model.Predict With Predictions(new_text_features)

not_hoax_percentage = probabilities of not hoax * 100 and round it for last 5
digit behind decimal
hoax_percentage = probabilities of hoax * 100 and round it for last 5 digit
behind decimal

probabilities_percentage = [not_hoax_percentage, hoax_percentage]

return predictions, probabilities_percentage
```

Gambar 10. Algoritma Data Fitur Sentimen

Sedangkan fitur *FactCheck* diperoleh dengan algoritma pada Gambar 11, yaitu dilakukan pemodelan *Multinomial Naïve Bayes* dengan data hasil *text preprocessing* terhadap label dan hasil pemodelannya digunakan pada algoritma pada Gambar 12 untuk menghitung probabilitas hoaks data fitur *FactCheck*.

```
Train After Text Preprocessing ()

train_text = load train text from database
train_label = load train label from database

texts_to_train=  preprocessing(train_text)

balanced_data = SMOTE(texts_to_train)

new_text_features =  TfIDF(balanced_data )
mnb = create MultinomialNaiveBayes object
mnb.fit(new_text_features, train_label)

store mnb in a binary file using pickle
```

Gambar 11. Algoritma Pemodelan Data *Text Preprocessing*

Predict Fact Check With Probability (string text)

trained_model = load trained model from file (FILE_FACT_CHECK_MODEL)

text to predict = preprocessing(teks)

new_text_features = TfIDF(new_text)
probabilities = trained_model.Predict With Probabilities(new_text_features)
predictions = trained_model.Predict With Predictions(new_text_features)

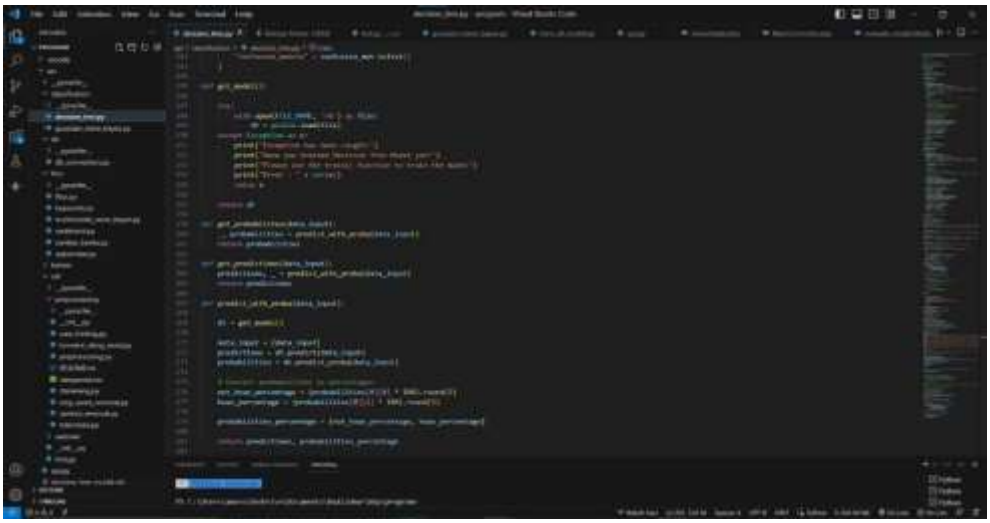
*not_hoax_percentage = probabilities of not hoax * 100 and round it for last 5 digit behind decimal*
*hoax_percentage = probabilities of hoax * 100 and round it for last 5 digit behind decimal*

probabilities_percentage = [not_hoax_percentage, hoax_percentage]

***return** predictions, probabilities_percentage*

Gambar 12. Algoritma Data Fitur *FackCheck*

Pengkodean *Python* dari algoritma untuk mendapatkan fitur-fitur *dataset* dilakukan dengan editor *VSCode* dengan tangkapan layar salah satu bagian ditunjukkan pada Gambar 13.



Gambar 13. Pengkodean dengan *VSCode*

Berdasarkan semua algoritma di atas maka sampel *dataset* yang diperoleh dari pemilihan dan kontruksi atribut data Mafindo ditunjukkan pada Gambar 14.

Keyword	Kapital	Sumber	Sentimen	FactCheck	Klasifikasi
0	4.093567261961441	NOT CREDIBLE	7.75922	100	Hoax
0	28.740740000120163	NOT CREDIBLE	61.8778	99.9886	Hoax
0	10.447761418435242	NOT CREDIBLE	51.1513	100	Hoax
0	0	CREDIBLE	69.5534	50	Valid
-1	7.737801969051361	NOT CREDIBLE	0.00001	100	Hoax
0	19.727891688357849	OTHERS	49.3178	99.9799	Hoax
0	8.457711338996887	OTHERS	16.8671	99.9885	Hoax
0	0.6000005868897166	NOT CREDIBLE	7.75922	100	Hoax
0	2.409638464450636	NOT CREDIBLE	87.3732	100	Hoax
0	18.292683362960815	NOT CREDIBLE	23.3328	98.7187	Hoax
0	0	NOT CREDIBLE	75.1756	100	Hoax
0	0	NOT CREDIBLE	69.5534	50	Valid
0	0	OTHERS	69.5534	50	Valid
0	14.049586653709412	NOT CREDIBLE	69.5534	90.9384	Hoax
0	2.9411764015449715	NOT CREDIBLE	9.75707	99.9798	Valid
0	3.375527262687683	NOT CREDIBLE	2.8024	100	Hoax
0	7.14285746216774	NOT CREDIBLE	69.5534	99.9752	Hoax
0	23.841059207916626	NOT CREDIBLE	72.7223	100	Hoax
0	9.48905125260353	NOT CREDIBLE	69.5534	17.0902	Hoax
0	21.052631735801697	NOT CREDIBLE	45.4688	99.9452	Hoax

Gambar 14. Sampel Dataset Penelitian

4.2. *Modeling*

Pemodelan *Decision Tree* pada data latih dilakukan dengan algoritma pada Gambar 15.

```
Train Model ()

all_train_scores = load all train scores from database
train_labels = load labels from database

balanced_train_scores, balanced_train_labels = SMOTE(all_train_scores,
train_labels )

dt = create DecisionTree object
dt.fit(balanced_train_scores, balanced_train_labels )

store dt in a binary file using pickle
```

Gambar 15. Algoritma Pemodelan

Setelah dilakukan pemodelan dilanjutkan dengan pengujian model dengan algoritma pada Gambar 16.

```
Predict With Probability (float score_capital, int score_keyword, string
score_sumber, float score_sentimen_negatif, float score_factcheck)

trained_model = load trained model from file
(FILE_HOAX_DETECTOR_MODEL)

all_scores = [score_capital, score_keyword, score_sumber,
score_sentimen_negatif, score_factcheck]


probabilities = trained_model.Predict With Probabilities(all_scores )
predictions  = trained_model.Predict With Predictions(all_scores )

not_hoax_percentage = probabilities of not hoax * 100 and round it for last 5
digit behind decimal
hoax_percentage = probabilities of hoax * 100 and round it for last 5 digit
behind decimal

probabilities_percentage = [not_hoax_percentage, hoax_percentage]

return predictions, probabilities_percentage
```

Gambar 16. Algoritma Pengujian

Algoritma pemodelan diproses dengan *Python* dengan semua *library* yang dibutuhkan. Beberapa *library* penting ditunjukkan hasil tangkapan layar pada Gambar 17.

```
# Library NLTK untuk preprocessing teks
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

# Untuk mengambil dataset dan melatih model
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

# Untuk perhitungan akurasi model
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score

# Untuk menyimpan model hasil train dalam bentuk file biner
import pickle

# Untuk mengambil dan menyimpan hasil skoring dataset di database MySQL
import mysql.connector
```

Gambar 17. *Library* yang Digunakan dengan *Python*

4.3. *Evaluation*

Metode *Confusion Matrix* [47] digunakan untuk menampilkan hasil pengujian. Dalam penelitian ini, *P* (*Positive*) digunakan sebagai hoaks dan *N* (*Negative*) sebagai valid (bukan hoaks), dengan demikian *TP* (*True Positive*) adalah kondisi hasil prediksi dan label sama-sama hoaks, *TN* (*True Negative*) adalah kondisi hasil prediksi dan label sama-sama valid, *FP* (*False*

Positive) adalah kondisi hasil prediksinya hoaks namun label berisi valid, *FN* (*False Negative*) adalah kondisi hasil prediksinya valid namun label berisi hoaks. Pengukuran dilakukan dengan *Accuracy* (Acc.=(TP+TN)/(TP+TN+FP+FN)) dan *F1 Score* (F1=(2*TP)/(2*TP+FP+FN))

Hasil pengujian terhadap *dataset* (*Set*) dengan jumlah data (N) yang berbeda-beda tanpa SMOTE [46] ditampilkan pada Tabel 3. Pengujian ini dilakukan juga pada data latih, tidak hanya pada data uji, sebagai percobaan untuk perbandingan.

Tabel 3. Hasil Pengujian Tanpa SMOTE

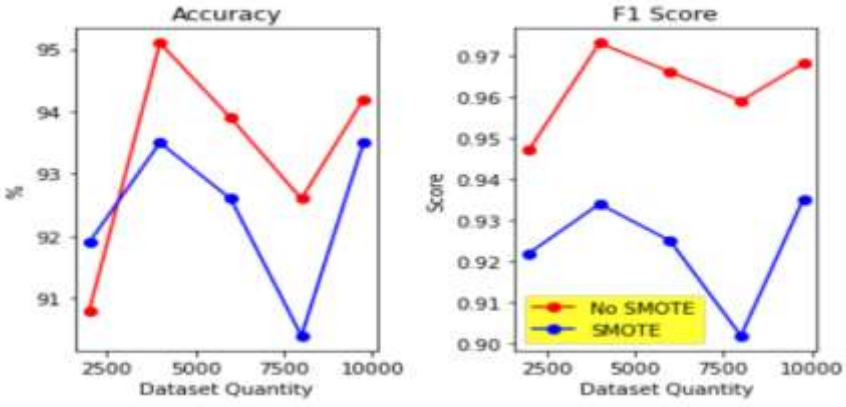
Set (N)	Pengujian		TP	FP	FN	TN	Acc.	F1
	Data	N						
9.756	Latih	7.804	7.180	0	32	664	99.6%	0,998
	Uji	1.952	1.744	58	56	94	94.2%	0,968
8.000	Latih	6.400	5.729	0	29	642	99.5%	0,997
	Uji	1.600	1.386	62	56	96	92.6%	0,959
6.000	Latih	4.800	4.290	0	30	480	99.4%	0,997
	Uji	1.200	1.034	27	46	93	93.9%	0,966
4.000	Latih	3.200	2.845	0	28	327	99.1%	0,995
	Uji	800	702	14	25	59	95.1%	0,973
2.000	Latih	1.600	1.421	0	21	158	98.7%	0,993
	Uji	400	333	12	25	30	90.8%	0,947

Sedangkan hasil pengujian pemodelan dengan menerapkan SMOTE (untuk menyeimbangkan jumlah data hoaks dan label) ada di Tabel 4.

Tabel 4. Hasil Pengujian dengan SMOTE

Set (N)	Pengujian		TP	FP	FN	TN	Acc.	F1
	Data	N						
9.756	Latih	7.804	7100	0	33	7171	99.8%	0,998
	Uji	1.952	1681	107	126	1662	93.5%	0,935
8.000	Latih	6.400	5738	0	33	5749	99.7%	0,997
	Uji	1.600	1275	123	154	1328	90.4%	0,902
6.000	Latih	4.800	4289	0	30	4321	99.7%	0,997
	Uji	1.200	993	72	88	1007	92.6%	0,925
4.000	Latih	3.200	2845	0	28	2878	99.5%	0,995
	Uji	800	667	43	51	679	93.5%	0,934
2.000	Latih	1.600	1395	0	20	1465	99.3%	0,993
	Uji	400	344	17	41	318	91.9%	0,922

Sebagai visualisasi evaluasi data uji, dari Tabel 3 dan 4 ditampilkan dalam bentuk grafik seperti pada Gambar 18.



Gambar 18. Visualisasi *Accuracy* dan *F1 Score*

4.4. *Deployment*

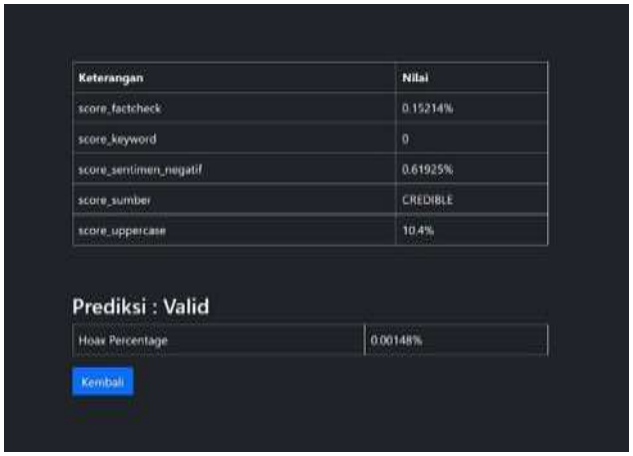
Hasil *deployment* purwarupa aplikasi berbasis *web* sederhana dengan memasukkan suatu berita baru untuk diprediksi hoaks atau valid, dengan cara *copy paste* ke dalam isian Konten Berita pada Gambar 19.

Gambar 19. Isian Konten Berita Baru

Selanjutnya untuk sumber berita bisa dipilih sesuai dengan sumber yang ada di data pelatihan, seperti pada Gambar 20, jika tidak ada yang sesuai maka dipilih sebagai data OTHERS untuk fitur Sumber.

Gambar 20. Pilihan Sumber Berita Baru

Selanjutnya tombol *Submit* ditekan untuk mendapatkan hasil prediksi dari model yang dibuat. Hasil prediksi disertakan informasi perhitungan data fitur hasil perhitungan model pada berita yang dimasukkan, seperti ditampilkan pada Gambar 21.



Gambar 21. Hasil Prediksi suatu Berita

Pengujian yang dilakukan terhadap data latih dan uji memberikan hasil akurasi dan F1 yang sudah tepat, karena pengujian pada data latih cenderung lebih baik namun dengan pemodelan *Decision Tree* tidak terjadi *overfit*. Perbedaan hasil pengujian data uji yang sedikit lebih rendah menunjukkan hasil pemodelan cukup baik.

Perbandingan pengujian pada kuntitas *dataset* yang berbeda-beda, tidak menunjukkan tren yang berarti pada nilai akurasi dan skor F1 karena terjadi naik turun tidak *significant*. Mengacu pada hasil visualisasi pada Gambar 18, maksimal selisih perbedaan akurasi tanpa SMOTE adalah 4,3% dan dengan SMOTE adalah 3.1%, sedangkan skor F1 maksimal selisih perbedaan tanpa SMOTE adalah 0.026 dan dengan SMOTE adalah 0,033.

Lebih lanjut pengaruh *imbalance dataset* dengan SMOTE, yaitu melakukan *oversampling* pada data yang berlabel lebih sedikit, secara umum memberikan kinerja yang sedikit lebih rendah dibandingkan tanpa SMOTE. Perbedaan selisih akurasi maksimal adalah 2,2% dan selisih F1 maksimal adalah 0,057, keduanya terjadi pada *dataset* 8000. Namun demikian model dengan SMOTE memiliki kecenderungan yang lebih stabil atau adil dalam mengklasifikasikan data, dengan perbandingan *False Negative* (FN) dan *False Positive* (FP) yang lebih seimbang dibandingkan model tanpa SMOTE, yang berarti tidak didominasi oleh label yang lebih banyak.

Pengaruh pemakaian fitur dari beberapa aspek dan diproses dengan sub pemodelan *Multinomial Naïve Bayes* pada fitur Sentimen dan *FactCheck* diduga memberikan dampak positif diperolehnya model yang stabil, namun perlu dilakukan penelitian lebih lanjut dalam hal ini.

5. Simpulan

Dalam penelitian ini, pemodelan *Decision Tree Classifier* pada data berita dengan fitur sumber berita, huruf kapital, kata kunci hoaks, sentimen negatif, *fact check* hoaks, dengan label klasifikasi valid dan hoaks, memberikan hasil yang memuaskan dengan nilai akurasi 93,5% dan skor F1 0,935 (angka tersebut secara kebetulan mirip). Penggunakan SMOTE dalam pemodelan *Decision Tree* untuk mengatasi *imbalance dataset* tidak terlalu berpengaruh pada hasil akurasi dan skor F1 namun memberi dampak kestabilan model terhadap kuantitas *dataset*.

Daftar Referensi

[1] Kumparan.com, “Dampak dari Pesatnya Perkembangan Teknologi di Era Digital.” Accessed: Jun. 02, 2023. [Online]. Available: https://kumparan.com/berita-update/dampak-dari-pesatnya-perkembangan-teknologi-di-era-digital-1vBkPOYNffj

[2] Sarwan, “Perspektif Hukum Pidana Mengenai Berita Hoaks.” Accessed: Jun. 02, 2023. [Online]. Available: https://www.kompasiana.com/inggamaulana45747/64785a338221996cf1383c52/perspektif-hukum-pidana-mengenai-berita-hoax-tentang-modus-pemerasan-pemotor-tabrakan-diri-kemobil-di-tangerang

[3] A. Bhattacherjee, “The effects of news source credibility and fact-checker credibility on users’ beliefs and intentions regarding online misinformation,” *Journal of Electronic Business & Digital Economics*, vol. 1, no. 1, pp. 24–33, Dec. 2022, doi: 10.1108/JEBDE-09-2022-0031.

[4] G. V. D. Kumar, M. V Jadhav, A. Tadisetti, and K. an, “A Deep Model on Hoax Detection Using Feed Forward Neural Network and LSTM,” *Webology*, vol. 17, no. 2, pp. 652–662, Dec. 2020, doi: 10.14704/WEB/V17I2/WEB17058.

[5] M. Zulfadhli, H. Hamdani, and L. Farokhah, “The Analysis of Hoax News Content on Facebook Reviewed from Theory of Critical Discourse Analysis and Linguistic Rules,” *Aksis : Jurnal Pendidikan Bahasa dan Sastra Indonesia*, vol. 5, no. 2, pp. 288–304, 2021, doi: 10.21009/aksis.050204.

[6] Utra T. Linge and A. F. Wicaksono, “Detection Of Negative Content (Hoax) On Microblog Data That Contains Covid-19 Information,” *Syntax Literate: Jurnal Ilmiah Indonesia*, vol. 7, no. 6, pp. 8820–8830, 2022.

[7] A. K. Darmawan, M. W. Al Wajieh, M. B. Setyawan, T. Yandi, and H. Hoiriyah, “Hoax News Analysis for the Indonesian National Capital Relocation Public Policy with the Support Vector Machine and Random Forest Algorithms,” *Journal of Information Systems and Informatics*, vol. 5, no. 1, pp. 150–173, Mar. 2023, doi: 10.51519/journalisi.v5i1.438.

[8] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. Rosal Ignatius Moses Setiadi, and R. S. Basuki, “Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization,” *TELKOMNIKA (Telecommunication Computing*

Electronics and Control), vol. 18, no. 2, pp. 799–806, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14744.

[9] D. Hidayat, A. Rohendi, D. Hanafy D, M. Christin, and N. Nur'aeni, "Fighting The Disinfodemic: Fact- Checking Management Of Hoax Covid-19 In Indonesia," *Profetik: Jurnal Komunikasi*, vol. 15, no. 2, pp. 272–286, Nov. 2022, doi: 10.14421/pjk.v15i2.1996.

[10] N. P. Satyawati, P. Utari, and S. Hastjarjo, "Fact Checking of Hoaxes by Masyarakat Antifitnah Indonesia," *International Journal of Multicultural and Multireligious Understanding*, vol. 6, no. 6, pp. 159–172, 2019.

[11] P.-M. Hui, C. Shao, A. Flammini, F. Menczer, and G. L. Ciampaglia, "The Hoaxy Misinformation and Fact-Checking Diffusion Network," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, pp. 528–530, Jun. 2018, doi: 10.1609/icwsm.v12i1.14986.

[12] G. Rebala, A. Ravi, and S. Churiwala, "Machine Learning Definition and Basics," in *An Introduction to Machine Learning*, Cham: Springer International Publishing, 2019, pp. 1–17. doi: 10.1007/978-3-030-15729-6_1.

[13] Potentia Analytics, "What Is Machine Learning: Definition, Types, Applications and Examples." Accessed: Jun. 10, 2023. [Online]. Available: https://www.potentiaco.com/what-is-machine-learning-definition-types-applications-and-examples/

[14] A. Y. Prayoga, A. I. Hadiana, and F. R. Umbara, "Deteksi Hoax pada Berita Online Bahasa Inggris Menggunakan Bernoulli Naïve Bayes dengan Ekstraksi Fitur Tf-Idf," *Jurnal Health Sains*, vol. 2, no. 10, pp. 1808–1823, 2021, doi: 10.46799/jsa.v2i10.327.

[15] A. Y. A. Nugraha and F. F. Abdulloh, "Optimasi Naive Bayes dan Cosine Similarity Menggunakan Particle Swarm Optimization Pada Klasifikasi Hoax Berbahasa Indonesia," *Jurnal Media Informatika Budidarma*, vol. 6, no. 3, pp. 1444–1451, 2022, doi: 10.30865/mib.v6i3.4170.

[16] H. Muhabatin, C. Prabowo, I. Ali, C. L. Rohmat, and D. R. Amalia, "Klasifikasi Berita Hoax Menggunakan Algoritma Naïve Bayes Berbasis PSO," *Informatics For Educators And Professional : Journal of Informatics*, vol. 5, no. 2, pp. 156–165, Jun. 2021, doi: 10.51211/itbi.v5i2.1531.

[17] E. Rasywir and A. Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin," *Jurnal Cybermatika*, vol. 3, no. 2, pp. 1–8, 2015.

[18] A. P. Kirana, G. B. Prasetyo, and E. W. Lestari, "The Detection of Indonesian Hoax Content about COVID-19 Vaccine using Naive Bayes Multinomial Method," *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 5, no. 1, pp. 13–19, Feb. 2023, doi: 10.35882/ijeeemi.v5i1.262.

[19] F. Prasetya and F. Ferdiansyah, "Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 4, no. 1, pp. 132–139, 2022, doi: 10.30865/json.v4i1.4852.

[20] H. Mustofa and A. A. Mahfudh, "Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes," *Walisongo Journal of Information Technology*, vol. 1, no. 1, pp. 1–12, Nov. 2019, doi: 10.21580/wjit.2019.1.1.3915.

[21] R. Wati, "Penerapan Algoritma Naive Bayes Dan Particle Swarm Optimization Untuk Klasifikasi Berita Hoax Pada Media Sosial," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 5, no. 2, pp. 9–14, 2020, doi: 10.33480/jitk.v5i2.1034.

[22] A. Kesumawati and A. K. Thalib, "Hoax classification with Term Frequency - Inverse Document Frequency using non-linear SVM and Naïve Bayes," *International Journal of Advances in Soft Computing and its Applications*, vol. 10, no. 3, pp. 115–128, 2018.

[23] F. Rahutomo, I. Y. R. Pratiwi, and D. M. Ramadhani, "Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia," *Jurnal Penelitian Komunikasi Dan Opini Publik*, vol. 23, no. 1, pp. 1–15, Jul. 2019, doi: 10.33299/jpkop.23.1.1805.

[24] A. Sudrajat, R. R. Wulandari, and E. Syafwan, "Indonesian Language Hoax News Classification Based on Naïve Bayes," *Journal of Applied Intelligent System*, vol. 7, no. 1, pp. 70–79, 2022, doi: 10.33633/jais.v7i1.5985.

[25] N. Agustina, A. Adrian, and M. Hermawati, "Implementasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Berita Palsu pada Sosial Media," *Faktor Exacta*, vol. 14, no. 4, pp. 206–213, 2021.

[26] S. Soleman, “Pemanfaatan Metode Klasifikasi Naïve Bayes Untuk Pendeteksi Berita Hoax Pada Artikel Berbahasa Indonesia,” *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, vol. 7, no. 2, pp. 83–93, 2021, doi: 10.24014/coreit.v7i2.14290.
[27] M. Z. Khan and O. H. Alhazmi, “Study and analysis of unreliable news based on content acquired using ensemble learning (prevalence of fake news on social media),” *International Journal of System Assurance Engineering and Management*, vol. 11, no. S2, pp. 145–153, Jul. 2020, doi: 10.1007/s13198-020-01016-4.
[28] R. R. Sani, Y. A. Pratiwi, S. Winarno, E. D. Udayanti, and F. Alzami, “Analisis Perbandingan Algoritma Naive Bayes Classifier dan Support Vector Machine untuk Klasifikasi Berita Hoax pada Berita Online Indonesia,” *Jurnal Masyarakat Informatika*, vol. 13, no. 2, pp. 85–98, 2022, doi: 10.14710/jmasif.13.2.47983.
[29] I. W. Santiyasa, G. P. A. Brahmantha, I. W. Supriana, I. G. G. A. Kadyanan, I. K. G. Suhartana, and I. B. M. Mahendra, “Identification Of Hoax Based On Text Mining Using K-Nearest Neighbor Method,” *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, vol. 10, no. 2, pp. 217–226, Jan. 2022, doi: 10.24843/JLK.2021.v10.i02.p04.
[30] E. Utami, A. F. Iskandar, W. Hidayat, A. B. Prasetyo, and A. D. Hartanto, “Covid-19 Hoax Detection Using KNN in Jaccard Space,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, pp. 255–264, 2021, doi: 10.22146/ijccs.67392.
[31] T. T. A. Putri, H. S. Warra, I. Y. Sitepu, M. Sihombing, and Silvi, “Analysis And Detection Of Hoax Contents In Indonesian News Based On Machine Learning,” *Journal Of Informatics Pelita Nusantara*, vol. 4, no. 1, pp. 19–26, 2019.
[32] B. Irena and E. B. Setiawan, “Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 711–716, Aug. 2020, doi: 10.29207/resti.v4i4.2125.
[33] N. A. Hasanah, N. Suciati, and D. Purwitasari, “Identifying Degree-of-Concern on COVID-19 topics with text classification of Twitters,” *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 7, no. 1, pp. 50–62, Feb. 2021, doi: 10.26594/register.v7i1.2234.
[34] C. W. Kencana, E. B. Setiawan, and I. Kurniawan, “Hoax Detection System on Twitter using Feed-Forward and Back-Propagation Neural Networks Classification Method,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 4, pp. 655–663, 2020, doi: 10.29207/resti.v4i4.2038.
[35] Mafindo.or.id, “MAFINDO – Masyarakat Anti Fitnah Indonesia.” Accessed: Jun. 11, 2023. [Online]. Available: https://www.mafindo.or.id/
[36] X. Wu *et al.*, “Top 10 algorithms in data mining,” *Knowl Inf Syst*, vol. 14, no. 1, pp. 1–37, 2008.
[37] T. Chandraveni, “CART (Classification And Regression Tree) in Machine Learning,” GeeksforGeeks. Accessed: Jul. 08, 2023. [Online]. Available: https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/
[38] N. Hotz, “What is CRISP DM? - Data Science Process Alliance.” Accessed: Jul. 01, 2023. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/
[39] Kementerian Ketenagakerjaan Republik Indonesia, “SKKNI Keahlian Artificial Intelligence (Data Science).” Accessed: Jul. 01, 2023. [Online]. Available: https://skkni.kemnaker.go.id/tentang-skkni/dokumen?area=data%20science&limit=20&page=1%20
[40] F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, “Foundations of JSON schema,” in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 263–273.
[41] Dewan Pers, “Data Perusahaan Pers,” Lembaga Dewan Pers. Accessed: Jul. 06, 2023. [Online]. Available: https://dewanpers.or.id/data/perusahaanpers
[42] R. Prakoso, “analisis-sentimen/kamus/positif_ta2.txt,” GitHub. Accessed: Jul. 05, 2023. [Online]. Available: https://github.com/ramaprakoso/analisis-sentimen/blob/master/kamus/positif_ta2.txt
[43] R. Prakoso, “analisis-sentimen/kamus/negatif_ta2.txt,” GitHub. Accessed: Jul. 05, 2023. [Online]. Available: https://github.com/ramaprakoso/analisis-sentimen/blob/master/kamus/negatif_ta2.txt
[44] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
[45] M. Pilgrim and S. Willison, *Dive Into Python 3*, vol. 2. Springer, 2009.

[46] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[47] M. K. Suryadewiansyah and T. E. E. Tju, “Naïve Bayes dan Confusion Matrix untuk Efisiensi Analisa Intrusion Detection System Alert,” *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 8, no. 2, pp. 81–88, 2022, doi: 10.25077/teknosi.v8i2.2022.81-88.