Klasifikasi Berita Palsu Berbahasa Inggris Menggunakan Metode *Random Forest* Dan *Naive Bayes*

Irin Lutfitasari[1], Ronggo Alit[2]

[1,2] Teknik Informatika, Fakultas Teknik, Universitas Negeri Surabaya
[1]irin.19086@mhs.unesa.ac.id
[2]ronggoalit@unesa.ac.id

***Abstrak*— Penyebaran informasi yang cepat dan tanpa penyaringan dapat mengakibatkan persepsi yang salah dan dampak negatif terhadap masyarakat. Harapan riset ini yaitu gna mengevaluasi serta membandingkan kinerja 2 metode klasifikasi, antaranya Random Forest serta Naïve Bayes, pada pengidentifikasian berita palsu. Dengan menggunakan empat skenario berbeda untuk setiap metode, penelitian ini memperoleh hasil akurasi yang signifikan. Metode Random Forest menunjukkan tingkat akurasi yang sangat tinggi, mencapai 99% dalam pengklasifikasian berita palsu. Sementara itu, Naïve Bayes mencapai tingkat akurasi sebesar 93%. Analisis hasil eksperimen menunjukkan bahwa Random Forest lebih efektif dalam menangani kompleksitas dan variasi dalam data berita palsu, yang tercermin dalam akurasi yang konsisten tinggi pada semua skenario. Sehingga, bisa ditarik hasil akhir Random Forest adalah metode yang lebih efisien dan andal dalam mendeteksi berita palsu dalam konteks penelitian ini. Temuan ini dapat menjadi dasar untuk pengembangan lebih lanjut dalam upaya menghadapi penyebaran berita palsu di era informasi digital.**

Kata Kunci*— Klasifikasi, *Machine Learning, Random Forest, Naïve Bayes, Python

I. PENDAHULUAN

Berita semakin mudah diakses seiring dengan berkembangnya teknologi dan dapat menyebar dengan luas. Berita yaitu himpunan informasi yang disampaikan mengenai suatu kejadian, opini, minat, situasi, kondisi/ulasan menarik yang sedang populer serta akan disebarluaskan ke publik. Berita sering kali disajikan kepada pihak ketiga/kelompok individu dengan media cetak, internet/dari mulut ke mulut [5].

Kemudahnya mengakses internet dan informasi membuat peluang masyarakat menyebar informasi di jejaring sosial maupun media online. Namun, penyebaran berita yang cepat tidak menjamin informasi yang kita terima bisa divalidasi dari sumber yang sah. Akibatnya, banyak berita palsu dari sumber yang tidak bisa dipercaya beredar di internet [9]. Distribusi informasi tersebut tidak terawasi, serta siapa pun bisa menyebarkan berita yang ambigu, bahkan berisi informasi palsu yang dikenal sebagai "hoaks" [5].

Berita palsu ialah informasi yang disampaikan dengan fakta yang salah/bohong. Penyebaran berita saat ini sangat cepat serta mudah diakses. Tentunya, jika penyebaran berita palsu terus berlanjut, hal ini bisa menimbulkan misinformasi yang diterima masyarakat [11]. Sebagai ilustrasi, terdapat sebagian masyarakat yang memanfaatkan platform media sosial untuk menyebarkan informasi. Tetapi, dari data yang tersedia, tidak semua individu memanfaatkan media sosial dengan penuh tanggung jawab pada penyebaran informasi. Dampak dari berita palsu itu sendiri bisa mengubah pandangan individu menjadi salah[7].

Informasi juga bisa dianggap sebagai alat politik yang validitasnya tidak bisa dijamin yang disebarkan dengan sengaja guna mencapai suatu sasaran tertentu. Informasi palsu juga memiliki potensi merugikan bagi individu/kelompok tertentu, serta bisa menimbulkan kepanikan di kalangan masyarakat yang berdampak negatif di kehidupan sehari-hari [14].

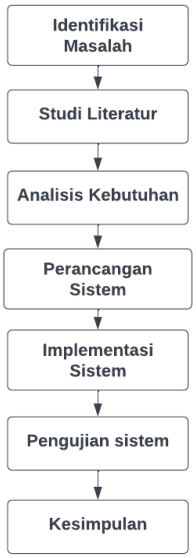
Banyaknya berita palsu setiap tahunnya yang menyebar di masyarakat. Usaha yang bisa dilaksanakan dengan mengklasifikasikan berita palsu dan berita benar menggunakan machine learning. Machine learning yaitu bagian dari kecerdasan buatan yang saat ini sedang berkembang pesat. Menggunakan *machine learning* diharapkan dapat mempersingkat waktu dan memiliki keakuratan tinggi. Sehingga diharapkan kedepanya masyarakat tidak lagi bingung dengan berita yang beredar.

Berlandaskan pemaparan tersebut, penulis melaksanakan riset akurasi presentasi klasifikasi berita palsu menggunakan metode random forest dan naïve bayes? Dan Metode manakah antara metode Random Forest dan Naïve Bayes yang memiliki nilai akurasi lebih tinggi dalam proses klasifikasi berita palsu?

II. METODOLOGI PENELITIAN

A. *Metode Penelitian*

Pada riset ini menerapakan langkah kuantitatif. Alur penelian yang akan dilaksanakan digambarkan sebagai berikut:



Gbr 1 Diagram Alur Metode Penelitian

Pada gambar 1 terdapat beberapa tahapan, pertama tahap identifikasi masalah dimana pada tahap tersebut permasalahan dalam penelitian ini adalah pengklasifikasian pada berita *online* berbahasa inggris. Penelitian ini menerapkan metode *data mining* klasifikasi sebagai rekomendasi kategori berita benar dan bohong. Penggunaan sistem ini untuk dapat membedakan kategori berita benar dan bohong tanpa dilakukan secara manual.

Setelah identifikasi masalah dilakukan, selanjutnya mencari studi literatur yang berfungsi sebagai referensi dimana pada tahap ini dilakukan pencarian teori-teori dan literatur relevan pendukung pada riset yang selaras pada tujuan riset ini dibuat. Penelitian ini menggali pengetahuan berhubungan dengan *machine learning, random forest, naïve bayes*, dan klasifikasi. Sumber lain penelitian ini meliputi buku, karya tulis ilmiah, dan jurnal.

Setelah informasi terkumpul, selanjutnya melakukan analisis kebutuhan. Pada tahap ini menganalisis kebutuhan yang dibutuhkan dalam penerapan proses pengkategorian atau klasifikasi dengan langkah *random forest* serta *naïve bayes* juga mengambil sumber data pada website *Kaggle.com.*

Setelah semua kebutuhan terkumpul, selanjutnya menyusun perancangan sistem. Pada tahap ini sistem dirancang serta dimanfaatkan pada riset dengan langkah *random forest* dan *naïve bayes* dalam klasifikasi berita online berbahasa inggris.

Setelah sistem sudah selesai dirancang, selanjutnya dilakukan penerapan sistem. Di langkah ini mengimplementasikan model sistem yang sudah dirancang sebelumnya yang telah dijelaskan pada latar belakang. Setelah dijalankan lalu dilakukan pengujian sistem.

Setelah pengimplementasian sistem dilakukan, selanjutnya akan dilakukan pengujian sistem dimana pada tahapan ini sistem diuji menggunakan dataset yang telah ditentukan.

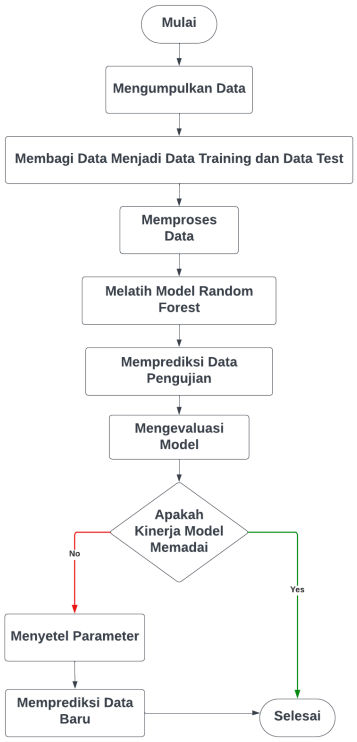
Setelah semua tahapan dilakukan maka akan diambil kesimpulan pada penelitian ini dimana peneliti menarik hasil akhir dari temuan penilian yang sudah dilaksanakan serta menjawab rumusan masalah dan tujuan dari penelian ini yang telah didefinisikan pada bab 1. Kemudian penulisan saran untuk pertimbangan penelitian selanjutnya.

B. Analisis Kebutuhan

1. Alat Penelitian
   a. Laptop Lenovo prosesor AMD A9
   b. Sistem operasi window 10 64-bit.
   c. Jupyter Notebook 6.0.3
   d. Python 3.8.3
   e. NumPy 1.18.5
2. Bahan riset
   a. Artikel terkait metode Randon Forest
   b. Artikel terkait metode Naïve Bayes
3. Dataset
   Di riset ini dataset di *website kaggle.com.* Dataset yang diambil yaitu berita palsu dan berita benar yang berjumlah 44.898 data berita.

C. Perancangan Sistem

1. *Preprocessing Data*
   Data yang telah diambil dari kaggle kedalam bahasa pemprograman *python. Preprocesing data* dilakukan sebelum menggunakan algoritma *machine learning.* Proses ini dilakukan agar untuk membersihkan, menformat, mempersiapkan data sehingga lebih sesuai untuk analisis.
   a. a. Pemurnian Data, merupakan proses memverifikasi data untuk menghilangkan entri yang hilang dan duplikat.
   b. b. Penanganan Data Kategorikal, adalah memberi label pada data dalam format kategorikal dengan menggunakan one hot encoding untuk memudahkan pemahaman oleh komputer.
   c. *c. Normalisasi, adalah proses transformasi data dengan menggunakan metode stopword.*
2. *Random Forest*
   Pohon keputusan (decision trees): pohon keputusan adalah struktur yang mebagi data bedasarkan serangkaian keputusan. Setiap simpul dalam pohon merepresentasikan keputusan bedasarkan fitur-fitur tertentu. Berikut adalah *flowchart Random Forest.*



Gbr 2 Flowchart Random Forest

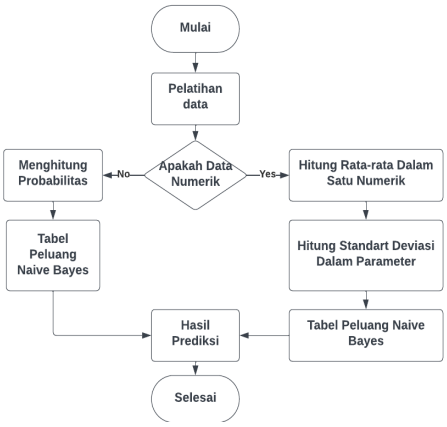
a. Mulai
b. Pengumpulan Data: Langkah pengumpulan Dataset sesuai klasifikasi
c. Membagi Data Menjadi Data Training dan Data Test: Pemisahan dataset menjadi bagian data pelatihan dan data pengujian
d. Memproses Data: Langkah-langkah seperti pengisian nilai yang hilang, mengubah variasi kategori menjadi format nomerik yaitu dalam rentan 0 atau 1.
e. Melatih Model Random Forest: Langkah pelatihan model random forest termasuk pembuatan banyak pohon keputusan dan proses bagging.
f. Memprediksi Data Pengujian: menggambarkan penggunaan model random forest untuk membuat prediksi pada dataset pengujian.
g. Mengevaluasi Model: evalusi kinerja model, termasuk penggunaan metrik seperti akurasi.

Apabila kinerja model memadai maka proses telah selesai, jika tidak memadai maka dilakukan proses.

h. Menyetel Parameter
i. Memprediksi Data Baru
j. Selesai

3. *Naïve Bayes*

Metode Naïve Bayes mengoperasikan dua jenis data untuk prediksinya, yaitu kumpulan data dan data evaluasi. Kumpulan data berperan sebagai sumber informasi untuk menghitung kemungkinan kejadian. Sementara data evaluasi berfungsi sebagai sampel untuk diprediksi berdasarkan kemungkinan yang telah dihitung tersebut.



Gbr 3 Flowchart Naïve Bayes

Informasi yang dapat diproses dalam Algoritma Naïve Bayes Classifier ada dua jenis, yaitu data kategorikal dan data numerik. Langkah-langkah pengolahan data tersebut dijelaskan di bawah ini:

a. Mulai
b. Baca data Training
c. Menghitung jumlah dan probabilitas, jika datanya numerik maka:
   1) Menyusun angka rerata atau rata-rata serta deviasi standar dari setiap parameter yang data-nya berbentuk angka.
      Rumus untuk menemukan nilai rerata (rata-rata) adalah seperti yang berikut ini:
      $$\mu = \varepsilon x n$$
      Dimana:
      - μ : nilai rata-rata hitung (*mean*) dari seluruh atribut
      - εx : jumlah nilai data x
      - n : banyaknya data

      Sementara rumus untuk memperoleh angka deviasi standar diformulasikan sebagai berikut:
      $$\sigma = \epsilon(x-\mu)^2 n$$
      Dimana:
      - σ : deviasi standar, menyatakan varian dari seluruh atribut
      - μ : *Mean* atau rata-rata dari seluruh atribut
      - n : banyaknya data
      - x : nilai dari suatu variabel x
   2) Mengalkulasikan probabilitas dengan menghitung jumlah informasi yang cocok dari jenis yang identik dibagi

dengan jumlah entri pada jenis yang bersangkutan.

d. Mendapatkan nilai dalam tabel *mean*, standar divisi dan probabilitas.
e. Menghasilkan prediksi.
f. Selesai.

i. Mengitung Akurasi

Akurasi merupakan penggambaran seberapa akurat model berdasarkan kedekatan tingkat nilai prediksi dengan nilai sebenarnya. Nilai akurat diperoleh melalui persamaan berikut ini:

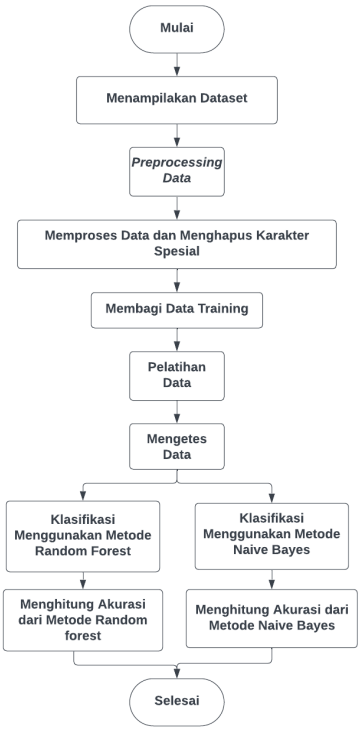
$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

Keterangan:

- TN = *True Negatif*
- TF = *True Positif*
- FP = *Fake Positif*
- FN = *Fake Negatif*

ii. Klasifikasi Menggunakan Metode *Random Forest* dan *Naïve Bayes*

Dalam penelitian ini diterapkan bebrapa proses dalam klasifikasi dengan algoritma *Random Forest* dan *Naïve Bayes*



Gbr 4 Flowchart Alur Klasifikasi

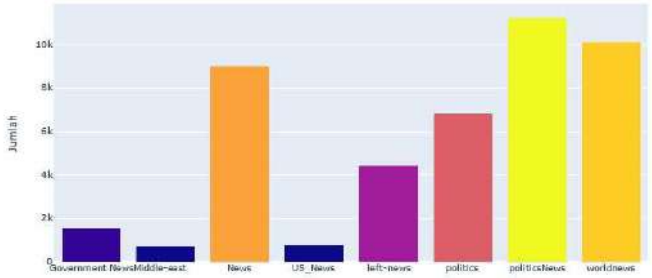
III. HASIL DAN PEMBAHASAN

A. *Deskripsi Data*

Dataset yang akan digunakan dalam penelitian ini berupa data yang diambil dari public dan sebanyak 44.898 data dalam bentuk file excel yang terdiri dari berita palsu dan berita benar dan memiliki 5 atribut diantaranya yaitu:

Title : Judul (Menjelaskan judul berita)
Text : Isi (Menjelaskan isi berita)
Subject : Tema (Menjelaskan tema berita)
Date : Tanggal (Menjelaskan tanggal terbit berita)
Real : Klasifikasi (Jika menunjukkan angka "0" maka berita tersebut dikategorikan palsu dan jika menunjukkan angka "1" maka berita tersebut dikategorikan benar)

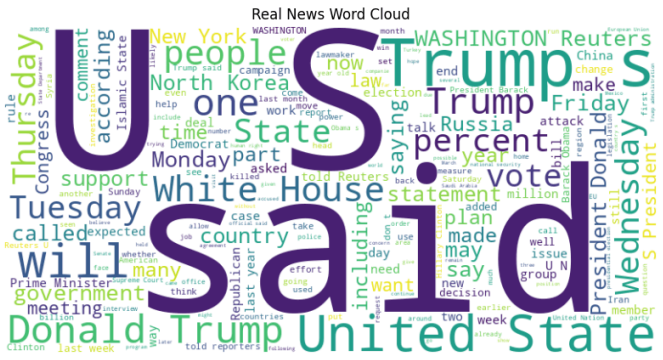
Data yang digunakan terdiri dari 47% berita benar dan 52% berita palsu. Data berita palsu memiliki data yang lebih besar daripada data berita benar. Data tersebut meliputi 21.417 data berita benar dan 23.481 berita palsu. Dari data tersebut terdapat beberapa kategori berita yaitu: Government News, Middle-east, News, US_News, left-news, politicsNews, Worldnews. Dengan jumlah data masing-masing Government News berjumlah 11.272 data, Middle-east berjumlah 778 data, News berjumlah 6.841 , US_News berjumlah 783 data , left-news 4.459 data, politicsNews 11.272 data, Worldnews 10.145 data. Seperti pada gambar dibawah ini.



Gbr 5 Kategori Data

Gambar diatas menjelaskan bahwa data yang digunakan terdapat jumlah data terbesar pada kategori politicsNews yang meliki lebih dari 10 ribu data berita, lalu diikuti worldnews dan data terkecil pada US_News dan Middle-east yang memiliki data kurang dari seribu data berita.

Untuk memberi gambaran tentang data yang digunakan, maka akan digunakan Wordcloud. Wordcloud digunakan pada tahap eksplorasi data dalam machine learning untuk memberikan visualisasi yang ringkas dan menarik dari data teks. Wordcloud secara visual menunjukkan frekuensi kata-kata yang paling umum dalam teks tersebut. Keuntungan menggunakan wordcloud adalah membantu pengguna dengan cepat mengidentifikasi kata-kata yang paling sering muncul dalam teks yang dianalisis. Wordcloud juga dapat membantu dalam pemahaman konten teks secara keseluruhan dan dapat digunakan sebagai alat awal untuk menemukan pola, tren, atau topik yang menonjol dalam dataset teks. Berikut adalah Wordcloud dari data yang digunakan.



Gbr 6 Wordcloud Berita Asli

Gbr 7 Wordcloud Berita Palsu

Dapat dilihat pada gambar wordcloud di atas, kata yang dicetak paling besar adalah kata yang paling sering muncul. Dalam hal ini, berita yang asli mengeluarkan kata "US" dan "SAID" lebih sering dari yang lain. Dan pada berita palsu, kata "ONES" dan "SAID" menjadi lebih sering daripada kata yang lain.

B. Implementasi

Pada tahapan ini terdapat 6 tahapan yaitu *preprocessing*, *cleaning data, Tokenizing, Deleting Stopwords, Lematization* dan *vectorization* menggunakan metode TF-IDF.

1. *Preprocessing*

Pada tahap *preprocessing* peneliti melakukan penghapusan secara otomatis data yang kosong dan mengkombinasikan data "*title*" dengan data "*text*" yang nantinya akan menghasilkan data "*combined text*") yang dapat dilihat pada gambar berikut.

```
[ ]  # Missing values (menghapus data yang kosong)
     df.dropna(inplace=True)
```

Gbr 8 Mengahpus Data Yang Kosong

```
[2]  # Combining 'text' and 'title' (menggabungkan title dan text hasilya combinasi text)
     df['combined_text'] = df['title'] + ' ' + df['text']
```

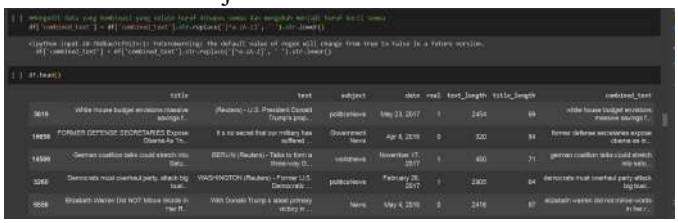
Gbr 9 Menggabungkan Data Title dan Text

	title	text	subject	date	real	text_length	title_length	combined_text
3619	White House budget envisions massive savings f...	(Reuters) - U.S. President Donald Trump's prop...	politicsNews	May 23, 2017	1	2454	69	White House budget envisions massive savings f...
16658	FORMER DEFENSE SECRETARIES Expose Obama As "In...	It is no secret that our military has suffered ...	Government News	Apr 8, 2016	0	320	84	FORMER DEFENSE SECRETARIES Expose Obama As "In...
14589	German coalition talks could stretch into Satu...	BERLIN (Reuters) - Talks to form a three-way G...	worldnews	November 17, 2017	1	450	71	German coalition talks could stretch into Satu...
5260	Democrats must overhaul party, attack big busi...	WASHINGTON (Reuters) - Former U.S. Democratic ...	politicsNews	February 26, 2017	1	2905	64	Democrats must overhaul party, attack big busi...
6556	Elizabeth Warren Did NOT Mince Words In Her R...	With Donald Trump's latest primary victory in ...	News	May 4, 2016	0	2416	87	Elizabeth Warren Did NOT Mince Words In Her R...

Gbr 10 Hasil Kombinasi Data

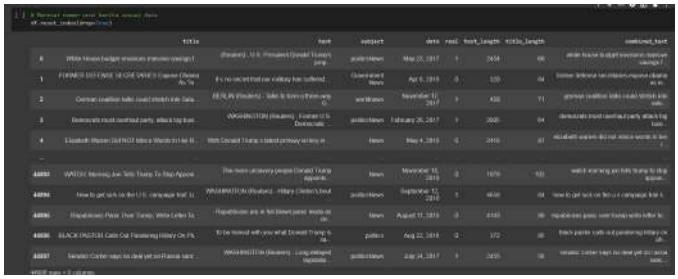
2. *Cleaning Data*

Pada tahap *cleaning data* peneliti melakukan edit data dengan cara mengedit "*combined data*" yang selain huruf akan dihapus dan semua huruf besar diubah menjadi huruf kecil semua.



Gbr 11 Perintah Mengedit data kombinasi dan hasil setelah dijalankan

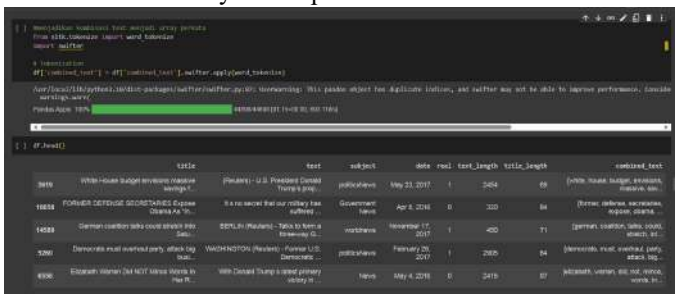
Setelah diedit isinya lalu peneliti melakukan reset nomor urut berita sesuai dengan data awal. Berikut hasil dari reset nomor urut berita.



Gbr 12 Perintah Reset Nomor Urut Berita dan Hasil Setelah Dijalankan

3. *Tokenizing*

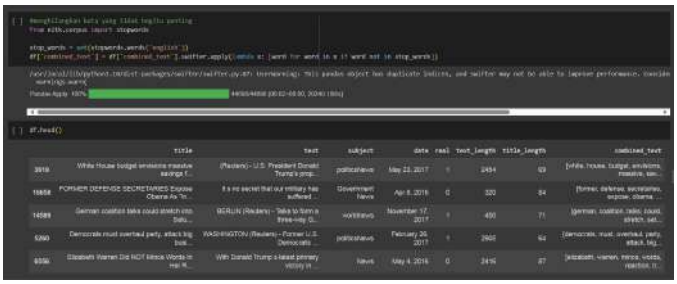
Pada tahap ini peneliti melakukan perubahan pada tipe data "*combined text*" yang akan diubah menjadi bentuk array dalam perkata.



Gbr 13 Perintah Merubah Kombinasi Text Menjadi Array Perkata dan Hasil Setelah Dijalankan

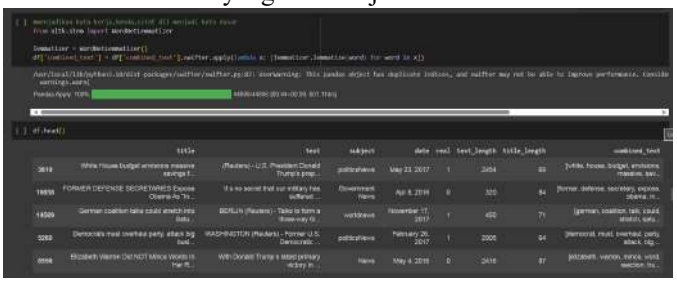
4. *Deleting Stopwords*

Pada tahap ini peneliti melakukan perubahan pada tipe data "*combined text*" dengan menghapus kata yang imbuhan yang tidak begitu penting.



Gbr 14 Perintah Menghapus Kata Yang Tidak Penting dan Hasil Setelah dijalankan

5. *Lematization*
   Pada tahap ini peneliti melakukan perubahan pada tipe data "*combined text*" yang akan mengubah semua kata yang ada menjadi kata dasar.



Gbr 15 Perintah Mengubah Semua Katan Menjadi Kata Dasar dan Hasil Setelah dijalankan

6. *Vectorization (TF-IDF)*
   Pada tahap ini peneliti mengonversi kumpulan teks menjadi representasi numerik berupa vektor. Disini peneliti menggunakan pendekatan metode *Term Frequency-Inverse Document Frequency* (TF-IDF) untuk menilai pentingnya suatu kata dalam dokumen secara relatif terhadap seluruh korpus teks. Dengan cara ini, setiap dokumen direpresentasikan sebagai vektor di dalam ruang fitur, di mana setiap dimensinya mewakili bobot kata berdasarkan frekuensinya dan dampaknya pada keseluruhan koleksi dokumen.

Gbr 16 Perintah vectorization (TF-IDF)

C. Evaluasi

Setelah tahap *preprocessing*, *cleaning data, Tokenizing, Deleting Stopwords, Lematization* dan *vectorization* selanjutnya dilakukan evaluasi dari hasil penelitian yang akan dibagi menjadi dua klasifikasi yaitu *Fake Classifier* dan *Subject Classifier.*

Setelah dibagi menjadi 2 klasifikasi peneliti juga membagi dalam 3 skenario data *test* dan data *training* pada masing-masing klasifikasi dengan perbandingan sebagai berikut:

1) 20% data *test* dan 80% data *training*
2) 25% data test dan 75% data training
3) 30% data test dan 70% data training
4) 35% data test dan 65% data training
5) 40% data *test* dan 60% data *training*

Data *train* dipergunakan untuk melatih metode yang digunakan dalam penelitian ini dan data test dipergunakan untuk menguji metode yang digunakan setelah melakukan training menggunakan data training. Berikut adalah rekap hasil evaluasi model dengan menggunakan *confusion matrix.*

1. *Fake Classifier*
   a. Skenario 1

Tabel 1 Skenario 1 (Fake Classifier)

Data *test* 20 % Data *train* 80%	*Random Forest*	*Naïve Bayes*
Accuracy	0.99	0.93
Precision	1.00	0.94
Recall	1.00	0.93
F1-Score	1.00	0.94

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 1 *Fake Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes.*

   b. Skenario 2

Tabel 2 Skenario 2 (Fake Classifier)

Data *test* 25 % Data *train* 75%	*Random Forest*	*Naïve Bayes*
Accuracy	0.99	0.93
Precision	1.00	0.93
Recall	1.00	0.93
F1-Score	1.00	0.93

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 2 *Fake Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes.*

   c. Skenario 3

Tabel 3 Skenario 3 (Fake Classifier)

Data *test* 30 % Data *train* 70%	*Random Forest*	*Naïve Bayes*
Accuracy	0.99	0.93
Precision	1.00	0.93
Recall	1.00	0.93
F1-Score	1.00	0.93

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 3 *Fake Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes.*

   d. Skenario 4

Tabel 4 Skenario 4 (Fake Classifier)

Data *test* 35 % Data *train* 65%	*Random Forest*	*Naïve Bayes*
Accuracy	0.99	0.93
Precision	1.00	0.93
Recall	1.00	0.93
F1-Score	1.00	0.93

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 4 *Fake Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes.*

e. Skenario 5

Tabel 5 Skenario 5 (Fake Classifier)

Data *test* 40 % Data *train* 60%	*Random Forest*	*Naïve Bayes*
Accuracy	0.99	0.93
Precision	1.00	0.93
Recall	1.00	0.93
F1-Score	1.00	0.93

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 5 *Fake Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes.*

Jika dilihat dari keseluruhan skenario, dapat dilihat bahwa konsistensi performa model Random Forest tetap terjaga walaupun proporsi data latih dan data uji berubah. Dalam setiap skenario, dari skenario 1 hingga skenario 5, Random Forest secara konsisten mencapai nilai yang tinggi di semua metrik - akurasi, presisi, recall, dan F1-Score. Hal ini menunjukkan bahwa model Random Forest memiliki *robustness* yang sangat baik dalam berbagai kondisi pembagian data.

Sebaliknya, model *Naïve Bayes*, meskipun menunjukkan performa yang cukup baik, tetap konsisten berada di bawah performa Random Forest dalam semua metrik evaluasi. Ini mengindikasikan bahwa walaupun *Naïve Bayes* adalah model yang cukup efisien dan mudah diimplementasikan, mungkin tidak sesuai untuk kasus yang kompleks atau di mana interaksi antar fitur sangat penting, seperti dalam kasus klasifikasi berita palsu.

2. *Subject Classifier*

a. Skenario 1

Tabel 6 Skenario 1 (Subject Classifier)

Data *test* 20 % Data *train* 80%	*Random Forest*	*Naïve Bayes*
Accuracy	0.71	0.70
Precision	0.42	0.51
Recall	0.43	0.45
F1-Score	0.42	0.45

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 1 *Subject Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes* pada *accuracy* data. Sedangkan pada *precision, recall,* dan *f1-score* metode *naïve bayes* menunjukan hasil yang lebih baik daripada metode *random forest*.

b. Skenario 2

Tabel 7 Skenario 2 (Subject Classifier)

Data *test* 25 % Data *train* 75%	*Random Forest*	*Naïve Bayes*
Accuracy	0.71	0.70
Precision	0.42	0.50
Recall	0.44	0.45
F1-Score	0.43	0.45

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 2 *Subject Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes* pada *accuracy* data. Sedangkan pada *precision, recall,* dan *f1-score* metode *naïve bayes* menunjukan hasil yang lebih baik daripada metode *random forest*.

c. Skenario 3

Tabel 8 Skenario 3 (Subject Classifier)

Data *test* 30 % Data *train* 70%	*Random Forest*	*Naïve Bayes*
Accuracy	0.71	0.69
Precision	0.43	0.50
Recall	0.45	0.45
F1-Score	0.44	0.45

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 3 *Subject Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes* pada *accuracy* data dan memiliki kesamaan hasil pada *recall* data. Sedangkan pada *precision* dan *f1-score* metode *naïve bayes* menunjukan hasil yang lebih baik daripada metode *random forest*.

d. Skenario 4

Tabel 9 Skenario 4 (Subject Classifier)

Data *test* 35 % Data *train* 65%	*Random Forest*	*Naïve Bayes*
Accuracy	0.72	0.70
Precision	0.45	0.51
Recall	0.46	0.46
F1-Score	0.45	0.45

Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 4 *Subject Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang

lebih baik daripada metode *naïve bayes* pada *accuracy* data dan memiliki kesamaan hasil pada *recall* data. Sedangkan pada *precision* dan *f1-score* metode *naïve bayes* menunjukan hasil yang lebih baik daripada metode *random forest*.

e. Skenario 5

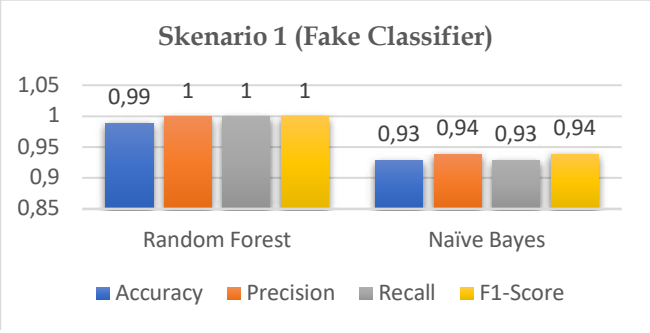
Tabel 10 Skenario 5 (Subject Classifier)

Data *test* 40 % Data *train* 60%	*Random Forest*	*Naïve Bayes*
Accuracy	0.73	0.70
Precision	0.45	0.49
Recall	0.47	0.45
F1-Score	0.46	0.45

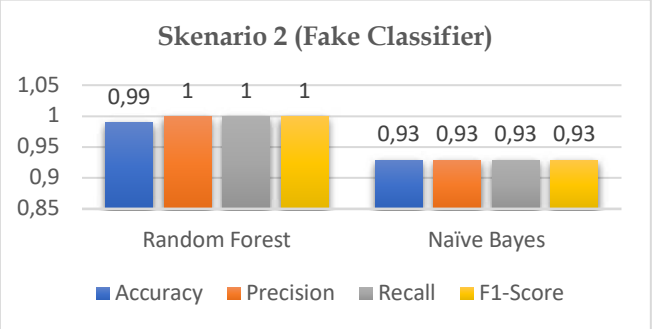
Hasil dari perbandingan antara metode *Random Forest* dengan metode *Naïve Bayes* pada skenario 5 *Subject Classifier*, kita dapat mengamati bahwa metode *random forest* mendapatkan hasil yang lebih baik daripada metode *naïve bayes* pada *accuracy, recall,* dan *f1-score* data. Sedangkan pada *precision* data metode *naïve bayes* menunjukan hasil yang lebih baik daripada metode *random forest*.

D. *Perbandingan Kinerja*

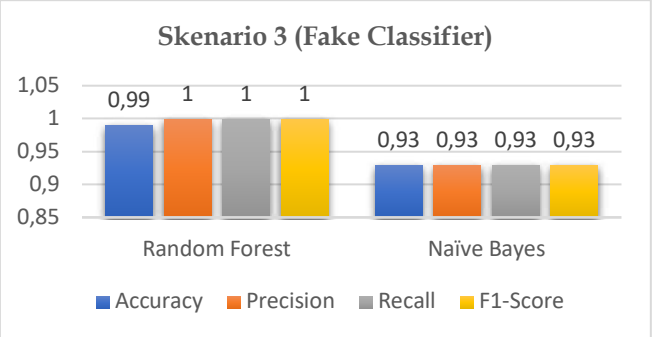
Pada tahap sebelumnya, dilakukan penelitian yang mendasarkan diri pada algoritma random forest dan naive bayes. Hasil penelitian menunjukkan bahwa algoritma random forest memberikan kinerja yang lebih baik dibandingkan dengan algoritma naive bayes pada beberapa tahap pengolahan data. Meskipun begitu, algoritma naive bayes tetap unggul dalam presisi data ketika digunakan sebagai subject classifier. Analisis lebih lanjut menunjukkan bahwa, meskipun random forest memberikan performa yang lebih baik secara umum, tetapi naive bayes tetap memiliki keunggulan tertentu terutama dalam konteks keakuratan pada klasifikasi subjek. Untuk memperoleh gambaran yang lebih jelas mengenai perbandingan hasil, dapat dilihat pada diagram batang yang disajikan di bawah ini. Diagram tersebut memberikan visualisasi yang mudah dipahami mengenai perbandingan kinerja kedua algoritma dalam berbagai aspek pengolahan data, menyoroti keunggulan masing-masing dalam konteks tertentu.



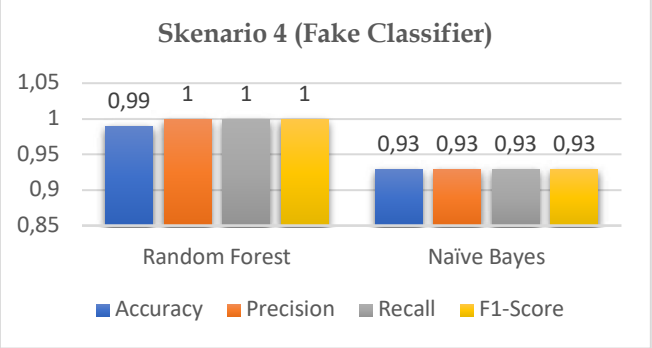
Gbr 17 Diagram Skenario 1 (Fake Classifier)



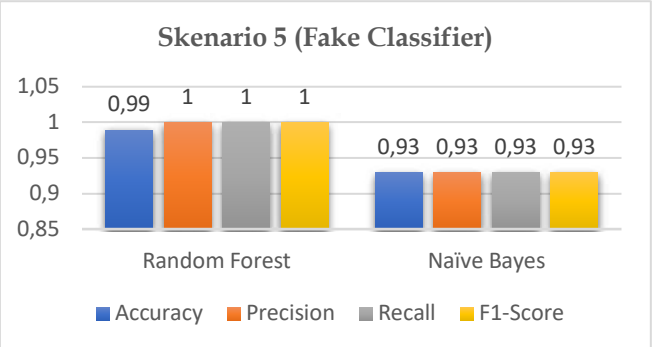
Gbr 18 Diagram Skenario 2 (Fake Classifier)



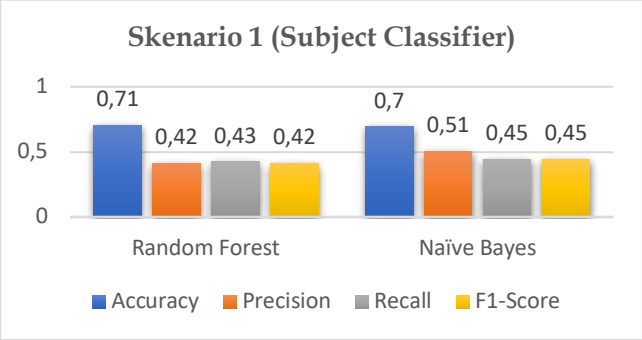
Gbr 19 Diagram Skenario 3 (Fake Classifier)



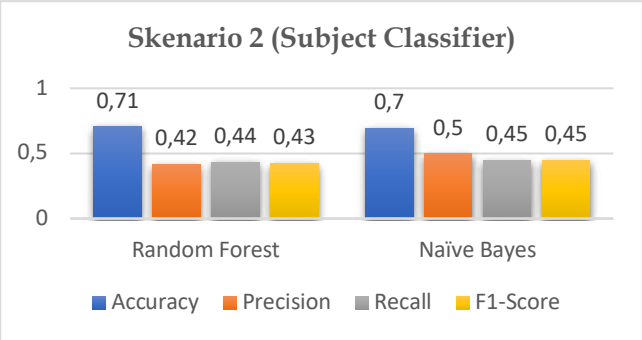
Gbr 20 Diagram Skenario 4 (Fake Classifier)



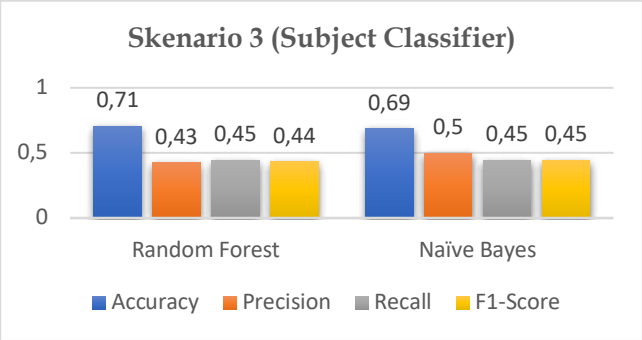
Gbr 21 Diagram Skenario 5 (Fake Classifier)



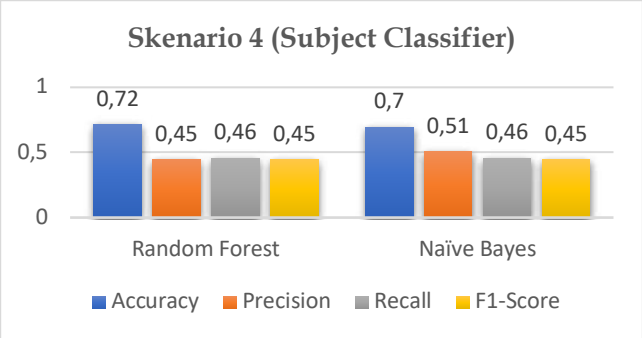
Gbr 22 Diagram Skenario 1 (Subject Classifier)



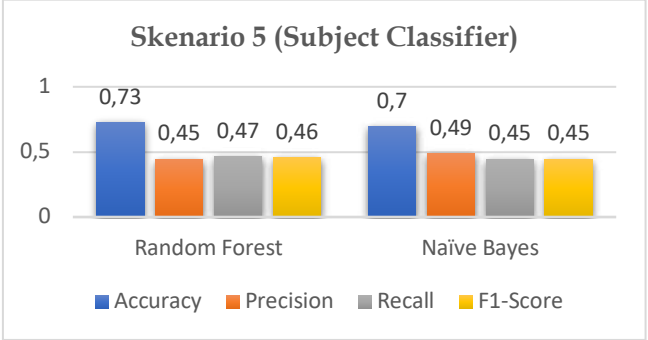
Gbr 23 Diagram Skenario 2 (Subject Classifier)



Gbr 24 Diagram Skenario 3 (Subject Classifier)



Gbr 25 Diagram Skenario 4 (Subject Classifier)



Gbr 26 Diagram Skenario 5 (Subject Classifier)

Salah satu alasan mengapa algoritma random forest mungkin memberikan kinerja lebih baik secara umum adalah karena kemampuannya untuk menangani sejumlah besar fitur dan memodelkan hubungan kompleks di antara variabel-variabel tersebut. Random forest merupakan ensemble learning algorithm yang menggabungkan hasil dari beberapa decision trees untuk menghasilkan prediksi yang lebih akurat. Kekuatan random forest terletak pada kemampuannya mengatasi overfitting dan memberikan generalisasi yang baik terhadap data yang belum pernah dilihat sebelumnya.

Sementara itu, keunggulan algoritma naive bayes sebagai subject classifier mungkin terkait dengan sifatnya yang bersifat sederhana dan efisien. Naive bayes bekerja berdasarkan asumsi bahwa setiap fitur dalam data adalah independen satu sama lain, yang meskipun sangat naif, dapat menghasilkan hasil yang baik terutama dalam konteks klasifikasi subjek. Keefisienan naive bayes membuatnya cocok untuk tugas-tugas di mana kompleksitas model bukanlah faktor kunci, dan di mana hubungan antara fitur-fitur mungkin relatif sederhana.

IV. KESIMPULAN

Berdasarkan riset di ranah keilmuan klasifikasi dengan menerapkan pendekatan random forest serta naive bayes, yang mengikuti perumusan tantangan, eksperimen, juga evaluasi kinerja data, bisa ditarik hasil akhir:

1. Perbandingan metode naïve bayes dan random forest pada dataset yang dimanfaatkan di riset ini memperlihatkan, langkah random forest memiliki akurasi lebih tinggi daripada metode naïve bayes dan metode random forest lebih efektif dalam mengklasifikasi berita yang palsu dengan jumlah data yang besar.
2. Kinerja algoritma metode random forest dinilai lebih akurat dalam skema fake classifier dengan menunjukan hasil sebesar 0,99 pada ketiga skenario yang diujikan. Dan dalam skema subject classifier metode random forest menunjukan hasil dalam range 0,71-0,73 pada ketiga skenario yang diujikan.

REFERENSI

[1] Arisandi, D., Indra, Z., & Kartini. (2021). Mengidentifikasi Pada Hasil Pencarian berita Online Dengan Teknik Web Scrapping dan Algoritma C4.5 .

[2] Cholissodin, I., Sutrisno, S., Hasanah, U., & Febiola, Y. (2019). *Buku Ajar AI , Machine Learning & Deep Learning* .

[3] Clinton, R., & Sengkey, R. (2019). Purwarupa Sitem Daftar Pelanggaran Lalulintas Berbasis Mini-Komputer Raspberry Pi. *Jurnal Teknik Elektro dan Komputer*.

[4] Hanifa, A., Fauzan, S. A., Hikal, M., & Ashfiya, M. (2021). PERBANDINGAN METODE LSTM DAN GRU (RNN) UNTUK KLASIFIKASI.

[5] Hendriyanto, M., & Betha, S. (2022). Penerapan Algoritma K-Neighbor Dalam Klasifikasi Judul Berita Hoak . *Jurnal Ilmiah Informatika (JIF)*.

[6] Iriadi, N., Priatno, & Ishaq, A. (2020). *Penerapan Data Mining dengan Rapid Miner.* Yogyakarta: Graha Ilmu.

[7] Krisna, D., & Salamah, U. (2022). PERBANDINGAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR UNTUK KLASIFIKASI BERITA HOAX KESEHATAN DI MEDIA SOSIAL TWITTER. *Jurnal Informatika Kaputama (JTIK)*.

[8] Muhabatin, H., Muhabatin, H., Ali, I., Rohmat, C., & Amalia, D. (2021). Klasifiasi berita hoax menggunakan naïve bayes berbasis PCO.

[9] Prasetya, F., & Ferdiansyah. (2022). Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes. *Jurnal Sistem Komputer dan Informatika (JSON)*.

[10] Prayoga, A., Hadiana, A., & Umbara, F. (2021). Deteksi Hoak Pada Berita Online Berbahasa Inggris Menggunakan Bernoullli Naïve Bayes Dengan Aktrasi Fitur TF-IDF. *Jurnal Syntax Admiration*.

[11] Ramadhan, N., Adhina, F., Segara, A., & Rakhmadani, D. (2022). Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression. *Kurnal Riset Komputer (JURIKOM)*.

[12] Berita palsu. (2018). Dalam *Wikipedia Bahasa Melayu, ensiklopedia bebas*. https://ms.wikipedia.org/w/index.php?title=Berita_palsu&oldid=4285086

[13] Mustofa, H., & Mahfudh, A. A. (2019). Klasifikasi Berita Hoax Dengan Menggunakan Metode Naive Bayes. *Walisongo Journal of Information Technology*, *1*(1), Article 1. https://doi.org/10.21580/wjit.2019.1.1.3915

[14] Rusdiana, I. (2018). Kognisi Pembaca Berita Palsu (Fake News) Di Media Online. *Kodifikasia*, *12*(2), 185. https://doi.org/10.21154/kodifikasia.v12i2.1520

[15] Zieniūtė, U. (2022, September 11). *Apa itu fake news atau berita hoaks? | NordVPN*. https://nordvpn.com/id/blog/berita-hoax/