# Indonesia Political Hoax News Classification Using IndoBERT


**Disusun Oleh:**

**NOVELLINA EDYAWATI**
**2702228223**
**SHERLY OKTAVIA WILLISA**
**2702262993**

**DTSC6008001 - TEXT MINING - 2025/2026**
**20 - 12 - 2025**
Github Repo

**SCHOOL OF COMPUTER SCIENCE**
**BINUS UNIVERSITY**

## 1. Project Summary

Projek ini bertujuan untuk mengembangkan sebuah model Deep Learning berbasis transformer yang dapat membedakan berita hoax dan berita valid berdasarkan konten teks secara kontekstual. Proses pengerjaan meliputi pengumpulan dan penyesuaian dataset, tahap preprocessing teks untuk menyiapkan data yang bersih, serta penerapan text representation melalui IndoBERT yang telah di pre-trained pada korpus bahasa indonesia. Model kemudian di fine-tune untuk tugas klasifikasi biner sehingga dapat mempelajari pola linguistik dan konteks yang khas pada berita politik. Hasil akhir proyek ini tidak hanya berupa model klasifikasi, tetapi juga dokumentasi alur analisis dan eksperimen yang terstruktur.

## 2. Problem Definition

Perkembangan media sosial mempercepat penyebaran suatu berita, khususnya berita politik. Namun dibalik itu, melalui media sosial, penyebaran berita palsu atau dikenal dengan HOAX juga semakin meningkat. Tidak sedikit pihak-pihak tidak bertanggung jawab mengeluarkan berita palsu yang menyudutkan pihak lain, dengan tujuan adu domba maupun merusak nama. Sebagai masyarakat yang pintar, tentu perlu mengolah dengan baik berita yang tersebar. Namun, seringkali proses verifikasi berita memakan waktu yang lama sehingga tidak mampu mengimbangi laju distribusi berita. Dengan memanfaatkan teknik text mining dikombinasikan dengan model Deep Learning berbasis transformer agar berita yang diterima dapat diperiksa dengan cepat apakah berita tersebut valid atau tidak. Melalui model klasifikasi berita hoax dan valid ini, diharapkan penyebaran berita hoax dapat diolah dengan baik oleh pembaca sehingga tidak merugikan orang lain.

## 3. Literature Review

Menurut KBBI, hoax adalah sebuah informasi bohong. Hooax dapat didefinisikan sebagai informasi yang sengaja disebarrkan untuk menyesatkan, menipu, membohongi pembaca dengan tujuan tertentu, seperti politik, ekonomi, atau hanya mencari perhatian publik semata. Hoax menjadi salah satu permasalahan sosial yang signifikan, khususnya di Indonesia. Seringkali, hoax menjadi penyebab terjadinya perselisihan antara beberapa pihak. Menurut Butar (2024), penyebaran hoax di media sosial memiliki dampak destruktif yang nyata terhadap masyarakat, termasuk memicu polarisasi dan ketidakpercayaan publik. Hal ini menegaskan urgensi pengembangan sistem yang mampu memberikan prediksi terkait berita yang tersebar, terutama dalam konteks isu-isu sensitif

seperti politik. Dalam penelitian yang dilakukan Hamdayani (2021) ditemukan beberapa fitur linguistik yang menjadi penanda kebohongan pada berita hoax, diantaranya berupa penggunaan kosa-kata atau diksi menggunakan ragam bahasa santai yang bersifat verba, nomina, dan adjektiva. Kemudian kebanyakan hoax menggunakan kalimat deklaratif (pernyataan), lalu imperatif (perintah), dan interogatif (pernyataan), untuk menekankan idelogis penulisnya. Selain itu, berdasarkan konteksnya teks hoax akan memacu konsumen teks untuk terus mengonfirmasi kebenaran isi teks yang dibaca dengan cara memberikan klarifikasi atau menunjukkan penelusuran dengan sumber yang tidak pasti. Yasa (2023) juga menyebutkan hoax memiliki ciri-ciri khusus diantaranya, tidak berdasarkan sumber yang jelas, baik dari fakta peristiwa yang disampaikan maupun fakta pendapat yang dimuat dalam berita, selain itu hoax dibuat dengan info atau cerita karangan yang berlebihan disertai dengan adanya ajakan untuk menyebarluaskan atau membuat berita itu menjadi viral, kebenaran dari berita hoax tidak dapat dibuktikan.

Beberapa penelitian sebelumnya telah mengeksplorasi penggunaan algoritma Machine Learning konvensional dalam mendeteksi hoax. Perkembangan algoritma klasifikasi tradisional menjadi landasan penting dalam studi terkait deteksi hoax di Indonesia melalui berbagai pendekatan eksperimental. Algoritma Naive Bayes sering digunakan sebagai tolak ukur awal karena efisiensinya dalam menangani data tekstual, seperti yang disampaikan oleh Rahutomo dkk. (2019) dalam penelitiannya. Mereka melakukan eksplorasi potensi dasar metode ini pada berita berbahasa Indonesia. Keandalan pendekatan berbasis probabilitas ini juga dipertegas oleh Ferdiansyah dkk. (2022) yang berhasil mengaplikasikannya secara efektif pada model klasifikasi hoax terkait COVID-19. Di sisi lain, penggunaan Support Vector Machine (SVM) mulai muncul sebagai alternatif yang lebih tangguh, seperti penelitian yang dilakukan oleh Sholeh dkk. (2023) dalam jurnalnya yang menyampaikan bahwa SVM cenderung memberikan akurasi yang lebih tinggi dibandingkan Naive Bayes karena kemampuannya yang lebih baik dalam menangani dimensi data yang tinggi. Selain pendekatan klasifikasi murni, terdapat pula eksplorasi dari sisi sistem pendukung keputusan melalui penggunaan metode Analytical Hierarchy Process (AHP). Sebagaimana diusulkan dalam penelitian oleh Kukuh dkk., algoritma ini memungkinkan penilaian kredibilitas berita dilakukan secara lebih struktural melalui metode perbandingan kriteria tertentu.

## 4. Data Collection

Dataset yang digunakan untuk projek ini adalah dataset yang kami temukan dari kaggle. Dataset ini merupakan kumpulan berita dengan topik politik yang ada di Indonesia. Kreator melakukan scrapping dari portal-portal berita terpercaya, yaitu CNN, Kompas, dan Tempo sebagai kelompok data valid. Sedangkan untuk data tidak valid (hoax) diambil dari situs Turnbackhoax yang merupakan situs pemeriksa fakta yang dikelola oleh MAFINDO (Masyarakat Anti Fitnah Indonesia) yang melakukan verifikasi dan pemeriksaan lebih lanjut terkait berita-berita yang sedang panas. Kreator

menyediakan dataset mentah dan yang sudah dibersihkan, disini kami menggunakan dataset cleaned yang telah disediakan. Pada folder dataset raw dan cleaner terbagi lagi untuk masing-masing platform berita. Untuk pengerjaan model kami, kami menggabungkan data cleaned dari seluruh platform.

Data terdiri dari 8-10 kolom, yang masing-masing memiliki deskripsi sebagai berikut.

a. Unnamed: 0 → indeks atau penomoran data didalamnya
b. Title → Judul berita
c. Timestamp → Tanggal berita di terbitkan
d. FullText → Seluruh teks pada artikel berita
e. Tag → keyword terkait berita
f. Author → Penulis berita
g. Url → Tautan untuk mengakses berita
h. Text_new → Isian dari berita secara lebih detail dan rapi
i. Hoax → Menandai apakah berita tersebut hoax atau tidak (0:valid, 1:hoax)
j. Politik (Turnbackhoax) → Menandai apakah berita tersebut membahas politik atau tidak (0:politik, 1: non-politik)
k. Narasi (Turnbackhoax) → Narasi yang terindikasi hoax
l. Clean narasi (Turnbackhoax) → Narasi yang telah dibersihkan (text cleaning)

## 5. Exploratory Data Analysis (EDA)

Kami melakukan beberapa tahapan untuk EDA untuk memahami struktur dan karakteristik dataset.

a. Struktur data

   Untuk memudahkan analisis struktur dataset, kami hanya menggunakan kolom news dan label.

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | news | 27447 non-null | object |
| 1 | label | 31353 non-null | int64 |

   Dimana pada kolom news terdapat sebanyak 27447 data dengan tipe data object, dan kolom label terdapat sebanyak 31353 data dengan tipe data integer.

b. Missing Values

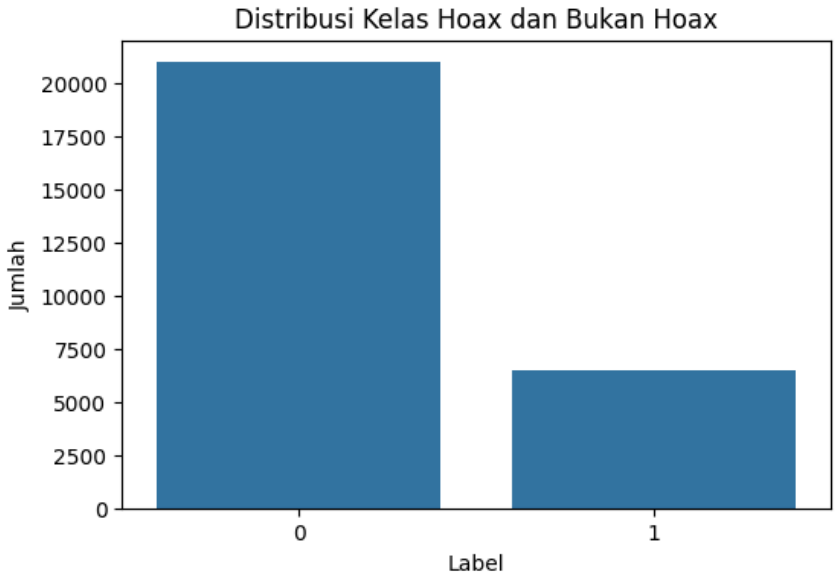
Ketika dilakukan pengecekan missing null values ditemukan sebanyak 3906 data news dengan nilai null, nantinya baris dengan nilai null ini akan dhilangkan pada tahap preprocessing.

Missing values:

| | |
|---|---|
| news | 3906 |
| label | 0 |

c. Duplicates Data
Ditemukan sebanyak 16 data duplikat untuk data news dan data label. Setelah dilakukan pembersihan data duplikat dan data null, maka jumlah data yang tersisa sebanyak 27431 data.

d. Distribusi kelas label



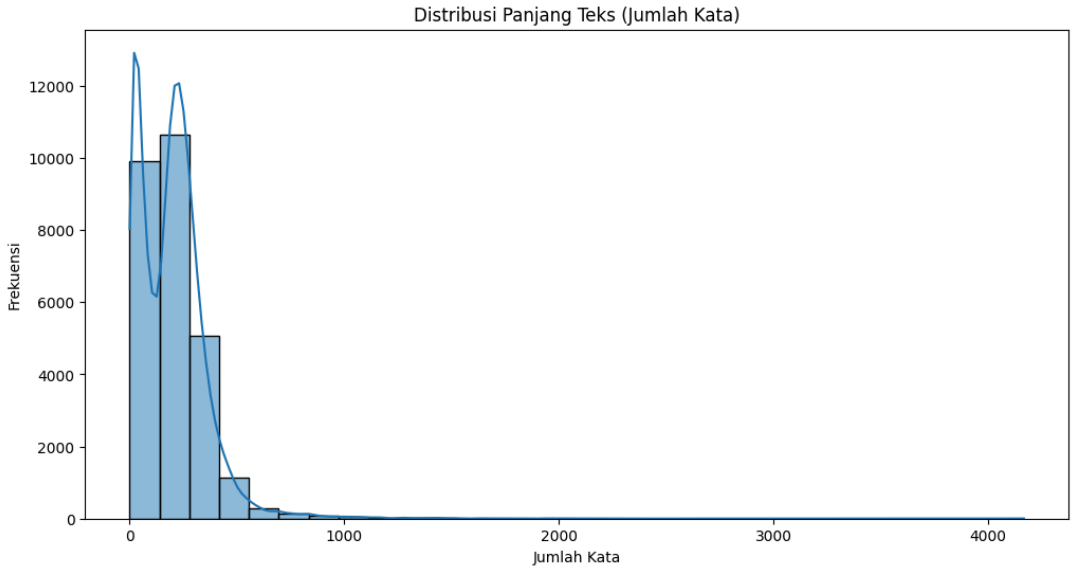

Distribusi kelas non-hoax (0) dan Hoax (1) menunjukkan adanya ketidakseimbangan data dimana data lebih didominasi oleh data berita valid

e. Statistik Deskriptif

```
Statistik Deskriptif Panjang Teks:
count    27431.000000
mean       202.695308
std        169.052520
min          1.000000
25%         71.000000
50%        198.000000
75%        279.000000
max       4167.000000
Name: text_length, dtype: float64
```

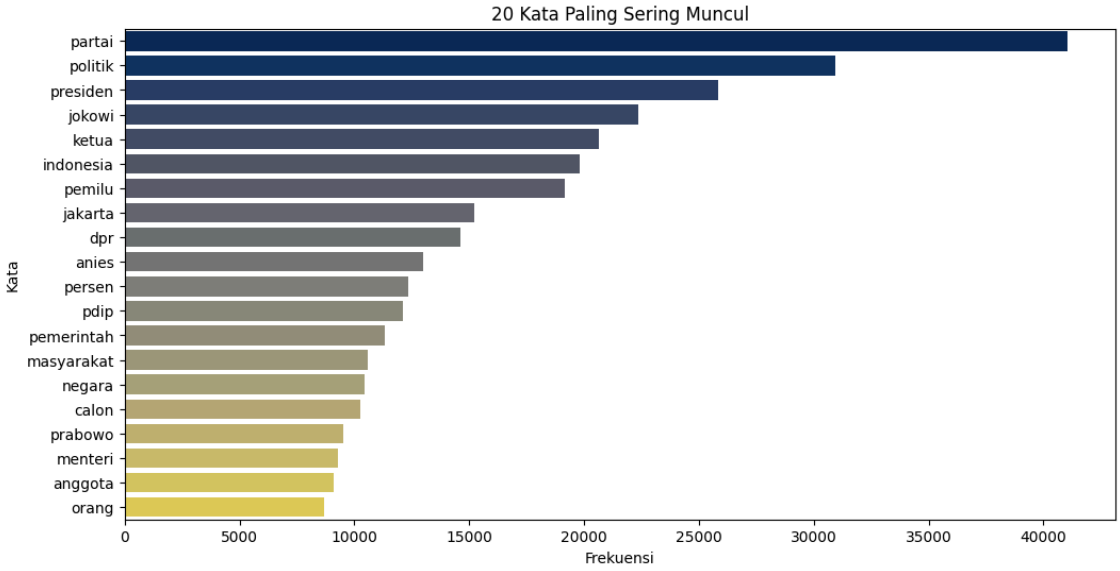
Rata-rata panjang teks berada di kisaran 203 kata, sehingga dapat disimpulkan sebagian teks dalam dataset memiliki panjang teks dengan golongan menengah tidak terlalu pendek, juga tidak terlalu panjang. Nilai median berada di kisaran 198 kata dimana nilainya berdekatan dengan nilai mean, hal ini menunjukkan data teks memiliki panjang yang konsisten berada di kisaran 190 kata hingga 200 kata. Standar deviasi pada teks sebanyak 169 kata menandakan variasi panjang teks yang cukup lebar, mengindikasikan adanya perbedaan signifikan antara teks pendek dengan teks yang sangat panjang. Kuartil 1 (25%) memiliki panjang kurang lebih 71 kata, menunjukkan sekitar 25% teks tergolong teks pendek, sedangkan untuk kuartil 3 (75%) memiliki panjang kurang lebih 279 kata, dapat disimpulkan sebanyak 75% teks memiliki panjang kurang lebih 279 kata. Dengan begtu dapat dikatakan mayoritas data teks berada pada rentang 71 kata hingga 279 kata.

f. Distribusi Panjang Teks Data



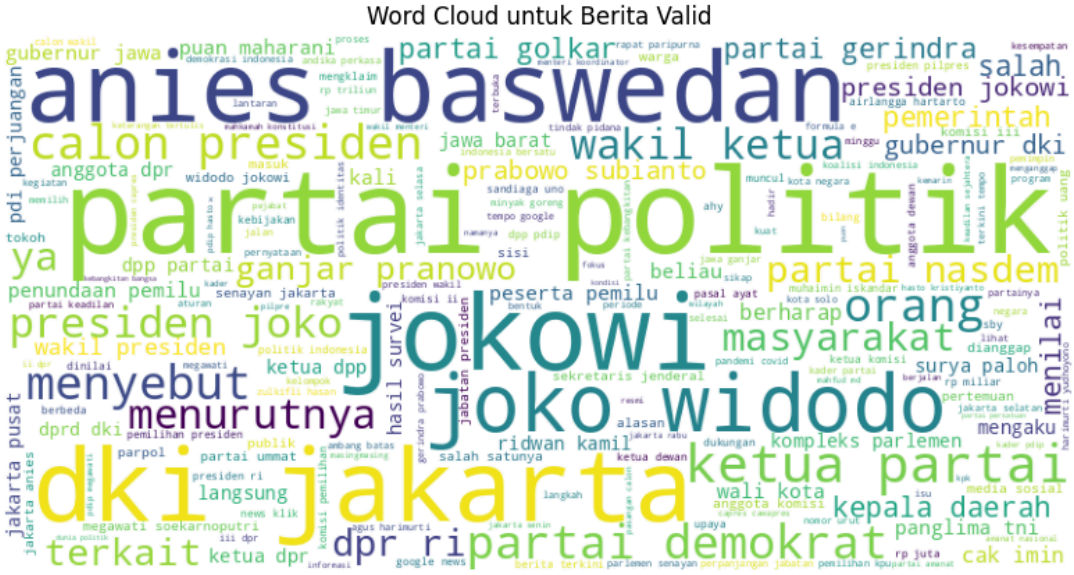

Berdasarkan histogram diatas, apabila kita amati distribusi panjang teks condong ke kiri atau right skewed, sehingga dapat dikatakan mayoritas dokumen memiliki jumlah kata yang relative sedikit, terlihat dari ekor yang melandai ke kanan. Puncak distribusi panjang kata berada di kisaran 300 kata hingga 500 kata, sehingga dapat disimpulkan sebagian besar dokumen memiliki jumlah kata sekitar 300 - 500 kata. Apabila kita lihat dengan seksama, ekor melandai dari angka 1000 ke kanan mencapai angka 4000, hal ini menandakan terdapat sedikit dokumen dengan panjang ekstrem dengan angka 1000 hingga 4000 kata.

g. Kata yang Paling Sering Muncul




Berdasarkan barplot ini, dapat kita lihat bahwa dataset ini didominasi oleh kata partai, politik, presiden, jokowi, ketua, indonesia, pemilu, jakarta, dpr, anies, dll. Dari plot ini kita dapat melihat dataset berita politik Indonesia didominasi dengan kata-kata yang berkaitan dengan tokoh dan posisi pemerintah di Indonesia.

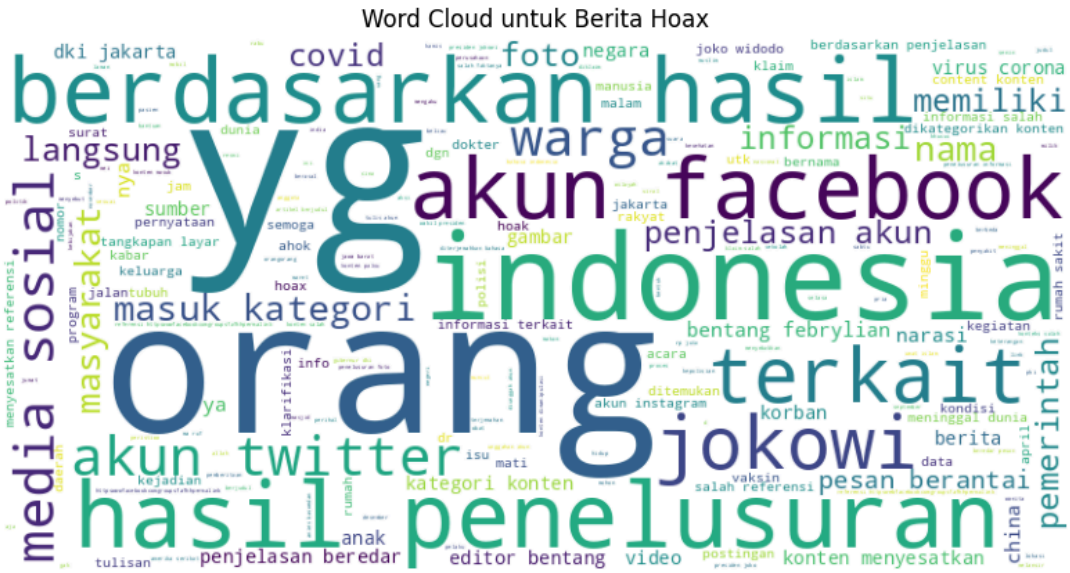
h. WordCloud untuk Berita Valid




Berdasarkan wordcloud berita valid, dapat kita amati wordcloud cenderung menonjolkan nama tokoh seperti Jokowi, Anies Baswedan, kemudian institusi resmi seperti kata ketua, wakil ketua, partai, dpr, golkar, gerindra, nasdem, pdi perjuangan. Selain itu apabila kita telusuri lebih dalam, berita valid cenderung menggunakan kata pelaporan yang bersifat jurnalistik seperti menurutnya, menyebut, terkait. Sehingga dapat kita simpulkan untuk label berita valid

memiliki pola yang menegaskan bahwa berita berorientasi pada fakta dan konteks formal, dengan bahasa jurnalistik yang netral.

i. WordCloud untuk Berita Hoax




Wordcloud berita hoax menunjukkan dominasi kata yang berkaitan dengan media sosial berita tidak resmi dan proses klarifikasi fakta yang merujuk pada ungahan individu bukan insititusi media. Dapat kita lihat kata-kata yang menonjol seperti akun twitter, akun facebook, berdasarkan hasil, hasil penelusuran, media sosial, orang, terkait, korban, anak, china, dll. Kata-kata tersebut menunjukkan bahwa informasi berasal dari unggahan indvidu pada media sosial seperti facebook dan twitter, bukan berasal dari institusi media berita resmi sehingga dapat disimpulkan bahwa berita merupakan narasi klaim yang tidak bersumber pada institusi resmi. Sehingga berita hoax biasanya akan memakai kata seperti hasil penelusuran, berdasarkan hasil untuk melakukan klarifikasi yang memperkuat argumen supaya mendapatkan kepercayaan publik yang membaca berita tersebut. Selain itu berita hoax juga menggunakan kata foto, video, postingan, korban dengan tujuan mencari perhatian dan memberi bukti palsu kepada publik agar mempercayai berita kebohongan tersebut.

## 6. Data Preprocessing

Data preprocessing dilakukan dengan tujuan untuk meningkatkan kualitas data teks sebelum digunakan pada proses analisis lanjutan dalam permodelan Natural Language Preprocessing (NLP) karena data berita ini berupa teks yang bersifat tidak terstruktur, Dibutuhkan beberapa tahapan preprocessing untuk menghilangkan noise dan bias,

mencegah label leakage, fokus model pada konteks dan makna berita. Tahapan data preprocessing ini meliputi:

- Konversi tipe data (awalnya object) menjadi tipe data string

```
s = str(s)
```

- Menghapus kata kunci yang menjadi indikator label hoax yang dapat berpotensi menyebabkan data leakage
- Menghapus struktur artikel klarifkasi fakta, menghindari model mempelajari format artikel yang dapat menyebabkan kebingungan model
- Menghapus referensi platform & sumber media, di dalam
- Menghapus istilah media sosial

```
patterns = [
    r'\(Narasi diterjemahkan.*?\)',
    r"\bhoax\b",
    r"\bsalah\b",
    r"\bhoaks\b",
    r"\bkeliru\b",
    r"\bmenyesatkan\b",
    r"\bmanipulasi\b",
    r"\bfitnah\b",
    r"\bhasut\b",
    r"\bvalid\b",
    r"\bmisinformasi\b",
    r"\bdisinformasi\b",
    r"\bfakta\b",
    r"\bklarifikasi\b",
    r"tidak\s*benar",
    r"berita\s*bohong",
    r"ini\s*salah",
    r"hasil periksa fakta",
    r"pemeriksa fakta",
    r"turnbackhoax",
    r"cek fakta",
    r"\btangkapan\s*layar\b",
    r"\bscreenshot\b",

    r'Penjelasan\s*&?\s*Fakta\s*:',
    r'Narasi\s*:',
    r'Klarifikasi\s*:',
    r'Fakta\s*:',
    r'Beredar\s*kabar\s*:',
    r"kategori\s*:",
    r"sumber\s*:",
    r"kesimpulan\s*:",
    r"penjelasan\s*:",
]
```

```
for pattern in patterns:
    s = re.sub(pattern, '', s, flags=re.IGNORECASE)
```

- Menghapus url & tautan

```
s = re.sub(r'http\S+|www\S+|https\S+', '', s)
```

- Manghapus tanda baca, simbol, dan melakukan lower case
- ```
  s = re.sub(r"[^a-zA-Z0-9\s]", " ", s).lower()
  ```
- Menghapus tag HTML

```
s = re.sub(r'<.*?>',' ', s)
```

- Normalisasi spasi

```
s = re.sub(r'\s+',' ', s).strip()
return s
```

Contoh teks sebelum dilakukan preprocessing:

Anies di Milad BKMT: Pengajian Menghasilkan Ibu-ibu Berpengetahuan Mantan Gubernur DKI Jakarta Anies Baswedan menghadiri acara tasyakuran Milad ke-42 tahun Badan Kontak Majelis Taklim (BKMT) di Istora Senayan, Jakarta, Selasa (21/2). Dia pun memuji eksistensi ibu-ibu pengajian mewujudkan keberhasilan pendidikan di dalam keluarga. Ia mengatakan selama 42 tahun usianya, BKMT telah menjadi teladan keberhasilan pendidikan dalam keluarga. BKMT, kata dia, menjadi bukti bahwa pengajian menghasilkan ibu-ibu yang lebih berpengetahuan. "BKMT menjadi bukti bahwa pengajian menghasilkan ibu-ibu yang lebih berpengetahuan. Ibu-ibu yang punya bekal untuk mendidik anak-anaknya, membuat rumah yang mencerminkan nilai Islam dan akhlak yang baik," kata Anies dikutip dari unggahan akun media sosial Instagram miliknya. Dalam unggahannya, Anies menyinggung soal BKMT yang tidak dapat dilepaskan dari sosok Tuty Alawiyah. Menurut Anies, Tuty bukan hanya seorang ustazah yang mampu memimpin ratusan Majelis Taklim hingga menjadi BKMT yang jangkauannya seluruh Indonesia, namun juga seorang ibu yang hebat dalam mendidik anak-anaknya. "Karena kami mengenal Prof Dailami Firdaus dan para saudaranya merupakan pribadi-pribadi yang berhasil di bidangnya, tak hanya itu mereka juga amat guyub serta saling support," tulis Anies. Dari 10 foto momen dirinya dalam Milad BKMT di Istora Senayan itu tampak pula hadir sejumlah tokoh politik dan pejabat negara. Beberapa di antaranya yang terlihat di foto dalam unggahan instagram Anies adalah Ketua Umum Partai Demokrat Agus Harimurti Yudhoyono (AHY), Ketum PAN yang juga Menteri Perdagangan Zulkifli Hasan, Wakil Ketua Dewan Pembina Gerindra yang juga Menteri Pariwisata dan Ekonomi Kreatif Sandiaga Uno, hingga Menteri BUMN Erick Thohir. "Alhamdulillah, dapat bersama-sama hadir dalam Tasyakur Milad ke-42 Badan Kontak Majelis Taklim (BKMT) di Istora Senayan," tulis Anies. (yoa/kid)

Contoh teks setelah dilakukan preprocessing:

anies di milad bkmt pengajian menghasilkan ibu ibu berpengetahuan mantan gubernur dki jakarta anies baswedan menghadiri acara tasyakuran milad ke 42 tahun badan kontak majelis taklim bkmt di istora senayan jakarta selasa 21 2 dia pun memuji eksistensi ibu ibu pengajian mewujudkan keberhasilan pendidikan di dalam keluarga ia mengatakan selama 42 tahun usianya bkmt telah menjadi teladan keberhasilan pendidikan dalam keluarga bkmt kata dia menjadi bukti bahwa pengajian menghasilkan ibu ibu yang lebih berpengetahuan bkmt menjadi bukti bahwa pengajian menghasilkan ibu ibu yang lebih berpengetahuan ibu ibu yang punya bekal untuk mendidik anak anaknya membuat rumah yang mencerminkan nilai islam dan akhlak yang baik kata anies dikutip dari akun media
```

sosial instagram miliknya dalam unggahannya anies menyinggung soal bkmt yang tidak dapat dilepaskan dari sosok tuty alawiyah menurut anies tuty bukan hanya seorang ustazah yang mampu memimpin ratusan majelis taklim hingga menjadi bkmt yang jangkauannya seluruh indonesia namun juga seorang ibu yang hebat dalam mendidik anak anaknya karena kami mengenal prof dailami firdaus dan para saudaranya merupakan pribadi pribadi yang berhasil di bidangnya tak hanya itu mereka juga amat guyub serta saling support tulis anies dari 10 foto momen dirinya dalam milad bkmt di istora senayan itu tampak pula hadir sejumlah tokoh politik dan pejabat negara beberapa di antaranya yang terlihat di foto dalam instagram anies adalah ketua umum partai demokrat agus harimurti yudhoyono ahy ketum pan yang juga menteri perdagangan zulkifli hasan wakil ketua dewan pembina gerindra yang juga menteri pariwisata dan ekonomi kreatif sandiaga uno hingga menteri bumn erick thohir alhamdulillah dapat bersama sama hadir dalam tasyakur milad ke 42 badan kontak majelis taklim bkmt di istora senayan tulis anies yoa kid

## 7. Text Representation

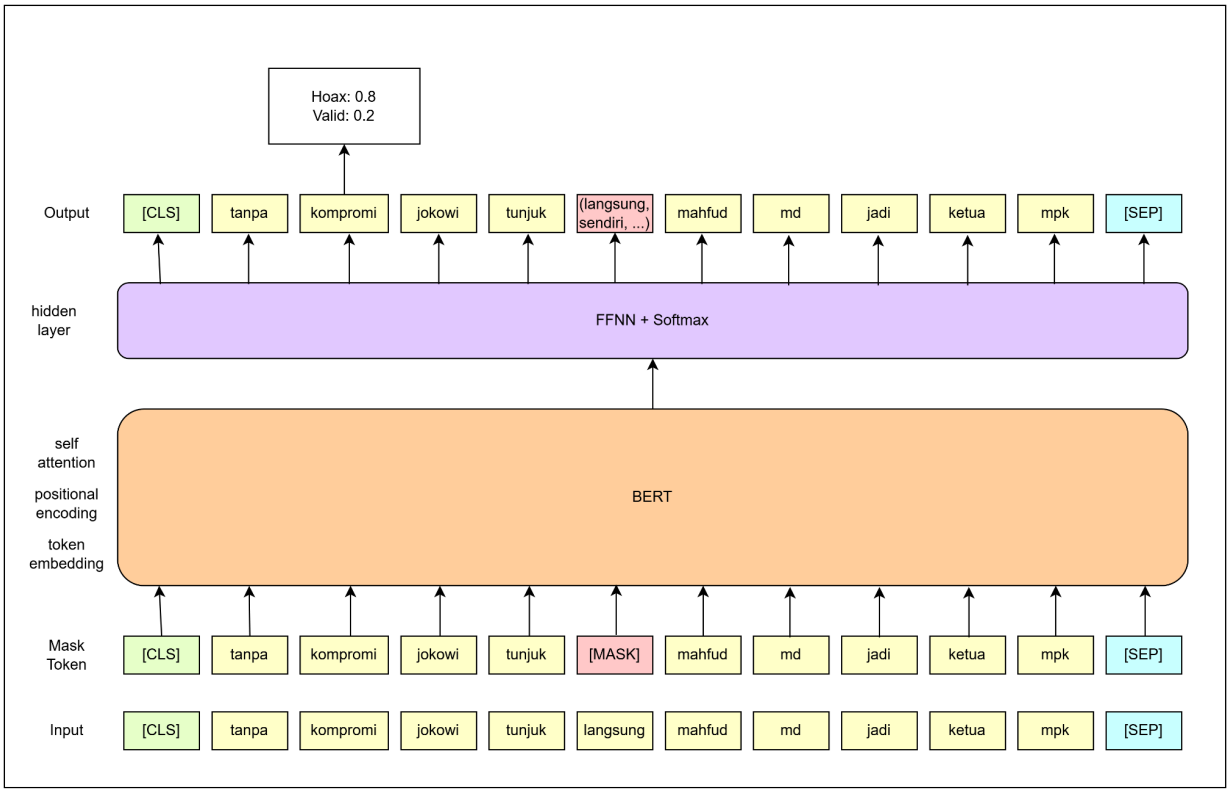
Representasi teks dilakukan menggunakan IndoBERT, yaitu model pre-trained berbasis transformer yang dikembangkan khusus untuk bahasa Indonesia. IndoBERT merepresentasikan teks dalam bentuk contextual embeddings, dimana setiap kata memiliki vektor representasi yang bergantung pada konteks kalimat secara keseluruhan. Dengan begitu, model mampu memahami makna kata secara lebih akurat, termasuk ambiguitas dan variasi penggunaan kata yang umum muncul dalam dokumen.

Sebelum diproses, dokumen berita akan di tokenisasi menggunakan IndoBERT tokenizer sehingga kata-kata kompleks atau tidak baku dapat dipecah menjadi subwords yang lebih representatif. Setiap token kemudian direpresentasikan sebagai kombinasi token embedding, position embedding, dan segment embedding. Text representation yang dihasilkan mampu menangkap informasi semantik global dari teks berita dan selanjutnya digunakan sebagai masukan pada lapisan klasifikasi untuk membedakan berita hoax dan non-hoax. Pendekatan text representation ini memungkinkan model mampu untuk menangkap pola bahasa, konteks, dan semantik yang kompleks sehingga meningkatkan performa klasifikasi.

## 8. Modeling

Tahap pemodelan pada proyek ini menggunakan IndoBERT sebagai model utama untuk melakukan klasifikasi berita politik ke dalam dua kelas, yaitu hoax dan valid. IndoBERT merupakan model pre-trained berbasis transformer yang telah dilatih pada korpus teks berbahasa Indonesia sehingga mampu memahami karakteristik linguistik dan konteks bahasa yang digunakan. Dalam pemilihan model dan tokenizer IndoBert yang kami gunakan kami memilih menggunakan model **indobert-base-p1** karena model pretrained ini merupakan model yang paling cocok dengan topik dan dataset yang kami gunakan, dimana dataset yang dilatihkan ke dalam model ini merupakan dataset yang

berasal dari berita Kompas, Tempo, dan Liputan6. Kami rasa dengan menggunakan model ini dataset yang kami gunakan (CNN, Tempo, Kompas, TurnbackHoax) akan menghasilkan performa yang baik karena model **indobert-base-p1** sendiri telah dilatih menggunakan dataset Kompas dan Tempo. Pada tahap ini, IndoBERT digunakan dengan pendekatan fine-tuning, dimana bobot model yang telah dilatih sebelumnya, disesuaikan kembali menggunakan dataset yang telah dikumpulkan sebelumnya untuk proyek ini agar lebih optimal.




IndoBert pada dasarnya merupakan model BERT, dimana di dalam model BERT input kalimat akan diubah ke dalam bentuk token kata. Kemudian, sebanyak 15% dari total kata akan dilakukan masking, hal ini berfungsi supaya model lebih memahami makna semantik pada suatu kalimat. Token CLS akan mejadi representasi secara keseluruhan untuk suatu kalimat. Sedangkan token SEP menjadi separator sebagai tanda pemisah token antar kalimat. Setelah dilakukan masking token kata mulai masuk ke dalam transformer BERT, dilakukan positional encoding, disini tokenizer indobert mengubah token kata menjadi bentuk list angka dimana angka tersebut merupakan bobot untuk setiap kata tergantung pada posisinya. Selain itu di dalam transformer BERT terdapat self attention, dimana bobot per kata akan dipengaruhi oleh bobot kata yang ada di depan dan di belakang kata tersebut. Setelah melewati proses encoder (transformer), informasi di proses pada hidden layer dengan menggunakan FFNN untuk menentukan

apa yang akan dilakukan terhadap informasi yang dihasilkan oleh mekanisme self attention. Dengan menggunakan softmax, informasi tersebut kemudian diubah ke dalam bentuk probabilitas, kemudian outputnya setiap token akan memiliki probabilitas berita hoax dan valid, probabilitas tertinggi akan menjadi kelas label untuk token kata tersebut.

Sebelum melakukan building model, kami membagi dataset kami menjadi data training, validation, dan testing dengan proporsi 80:10:10. Sehingga didapatkan data training sebanyak 21944 data, data validation sebanyak 2743 data, dan data testing sebanyak 2744 data. Pertama-tama kami melakukan training model dengan base line model menggunakan parameter

- Epoch : 5
- Maximal length token: 128
- Learning rate: 2e-5
- Batch Size: 16

```
Training using: cuda
Epoch 1, Loss: 0.0002372550661675632
Training using: cuda
Epoch 2, Loss: 0.00475166505202651
Training using: cuda
Epoch 3, Loss: 0.000643590756226331
Training using: cuda
Epoch 4, Loss: 0.1816318929195404
Training using: cuda
Epoch 5, Loss: 5.045403668191284e-05
```

Berdasarkan hasil training dengan base line model ini kami memperoleh hasil loss yang cukup rendah dan akurasi yang baik dalam mengklasifikasikan berita hoax dan valid. Namun ketika melakukan evaluasi model baseline, kami melihat adanya imbalance data, sehingga kami mencoba melakukan handling imbalance data dengan menggunakan 2 metode, Weighted Random Sampler dan melakukan cut data dengan kelas mayoritas supaya jumlahnya sama dengan kelas minoritas.

Selain itu kami juga melakukan hyperparameter tuning untuk mendapatkan performa terbaik model. Kami mencoba melakukan pencarian hyperparameter dengan parameter grid berikut:

- Learning rate : [2e-5, 3e-5]
- Batch Size : [16, 32]
- Epochs : [3, 5]

WeightedRandomSampler merupakan salah satu metode yang dapat digunakan untuk menghandle imbalance data khususnya untuk data teks. Cara kerja WRS ini dengan menciptakan bobot untuk setiap datanya, dengan menggunakan WRS data dari kelas minoritas akan memiliki bobot yang lebih besar daripada bobot kelas mayoritas.

Sehingga bobot yang lebih besar akan lebih sering muncul pada sampling. Dengan metode ini, kami memperoleh sebanyak 10909 data train dengan kelas valid (0), dan 11035 data train dengan kelas hoax (1).

$$Weight = \frac{1}{Jumlah\ data\ pada\ kelas\ x}$$

Setelah dilakukan pencarian hyperparameter tuning diperoleh, best parameter untuk model ini yaitu:

- Learning rate : 2e-5
- Batch Size :32
- Epochs : 5

Setelah best parameter hasil parameter tuning ditentukan, diperoleh hasil training dengan Validation F1–score dengan nilai yang lebih tinggi dibandingkan dengan model baseline.

```
Training | LR=2e-05, BS=32, Epochs=5
Epoch 1/5 | Loss: 0.0481
Epoch 2/5 | Loss: 0.0163
Epoch 3/5 | Loss: 0.0078
Epoch 4/5 | Loss: 0.0059
Epoch 5/5 | Loss: 0.0056
Validation F1: 0.9923
```

Kami juga mencoba melakukan training serta evaluation menggunakan data train yang kami cut jumlahnya menyesuaikan dengan jumlah data minoritasnya untuk memastikan ulang tidak ada data leakage. Dengan melakukan cutting data, diperoleh kelas data training untuk kelas valid (0) sebanyak 5189 data, dan kelas data hoax (1) sebanyak 5188 data. Dimana hasil training yang dihasilkan cukup baik, meskipun terdapat selisih sedikit dengan hasil training dengan dataset imbalance.

```
Epoch 1, Loss: 0.03165602311491966
Epoch 2, Loss: 0.00017230160301551223
Epoch 3, Loss: 0.00015433758380822837
Epoch 4, Loss: 0.0012689788127318025
Epoch 5, Loss: 6.691326416330412e-05
```

Metode validasi yang digunakan hold-out validation, dimana metode ini membagi dataset menjadi data training, validation, dan testing. Data validation digunakan untuk memantau performa selama proses training untuk mendeteksi adanya overfitting.

## 9. Evaluation

Pada tahap evaluasi, kami menggunakan 4 metrik evaluasi untuk klasifikasi yaitu *accuracy, precision, recall, dan f1-score* untuk mengukur sejauh mana model mampu

melakukan klasifikasi berita hoax dan valid secara akurat. Dengan menggunakan keempat metrik evaluasi tersebut diharapkan dapat menunjukkan pemahaman model dalam membedakan kelas valid dan hoax.

Accuracy mengukur proporsi prediksi yang benar terhadap seluruh data uji, dengan kata lain akurasi memberikan gambaran performa model secara general, namun akurasi tidak bisa dijadikan sebagai patokan utama khususnya ketika adanya ketidakseimbangan pada dataset.

Precision mengukur seberapa tepat model dalam melakukan prediksi suatu kelas positif, dengan kata lain precision merupakan perbandingan prediksi kelas yang ditebak secara tepat dari seluruh kelas positif yang diprediksi.

Recall mengukur sebarapa tepat model menemukan seluruh data yang benar-benar tepat termasuk ke dalam kelas positif. F1-score merupakan rata-rata dari precision dan recall, biasanya metrik ini yang dijadikan sebagai acuan utama performa model karena memberikan keseimbangan antara ketepatan dan kelengkapan prediksi yang dihasilkan model, f1-score akan sangat berpengaruh jika kondisi data imbalance.

Berdasarkan hasil pengujian pada data uji dengan menggunakan baseline model yang tidak menggunakan hyperparameter tuning dan tidak dilakukan penyeimbangan data.

```
Mulai Evaluasi...
Accuracy : 0.9898
F1 Score : 0.9897

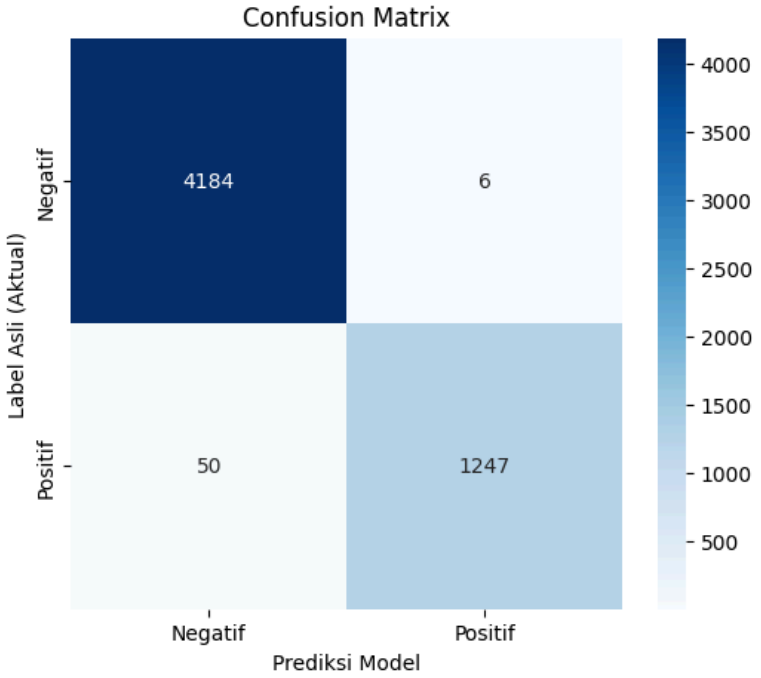
Detail Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      4190
           1       1.00      0.96      0.98      1297

    accuracy                           0.99      5487
   macro avg       0.99      0.98      0.99      5487
weighted avg       0.99      0.99      0.99      5487
```

Diperoleh akurasi sebesar 0.99, menandakan hampir seluruh data pada kelas hoax dan valid berhasil diprediksi secara benar dari seluruh data uji. Precision model mencapai angka 0.99 untuk kelas 0 menandakan hanya sekitar 1 persen data yang gagal diprediksi secara tepat sebagai kelas valid (0). Sedangkan, untuk kelas hoax (1) precision nya mencapai nilai 1.0 menandakan seluruh data yang diprediksi sebagai hoax, benar-benar data yang ada pada kelas hoax. Recall kelas valid (0) mencapai 1.0 menandakan model berhasil menangkap seluruh data valid yang tepat. Sedangkan untuk kelas hoax (1) memiliki recall sebesar 0.96, mengindikasikan sebanyak 4% data hoax lolos dimisklasifikasi sebagai data valid. F1-score kedua kelas mencapai nilai 0.99 dan 0.98

menandakan model sudah cukup baik membedakan kelas valid dan hoax meskipun terdapat sedikit sekali misklasifikasi pada kelas hoax yang diklasifikasikan sebagai valid.




Confusion matrix untuk data yang dujikan pada baseline model yang tidak dilakukan penyeimbangan data menunjukkan bahwa sebanyak 4184 data kelas valid berhasil diklasifikasikan sebagai kelas valid, hanya terdapat 6 data valid yang salah diklasifkasikan sebagai kelas hoax. Sedangkan untuk kelas hoax terdapat 1247 data yang berhasil diprediksi secara tepat, dan sebanyak 50 gagal diprediksi secara tepat (4% data kelas hoax yang lolos dimisklasifikasi sebagai kelas valid).

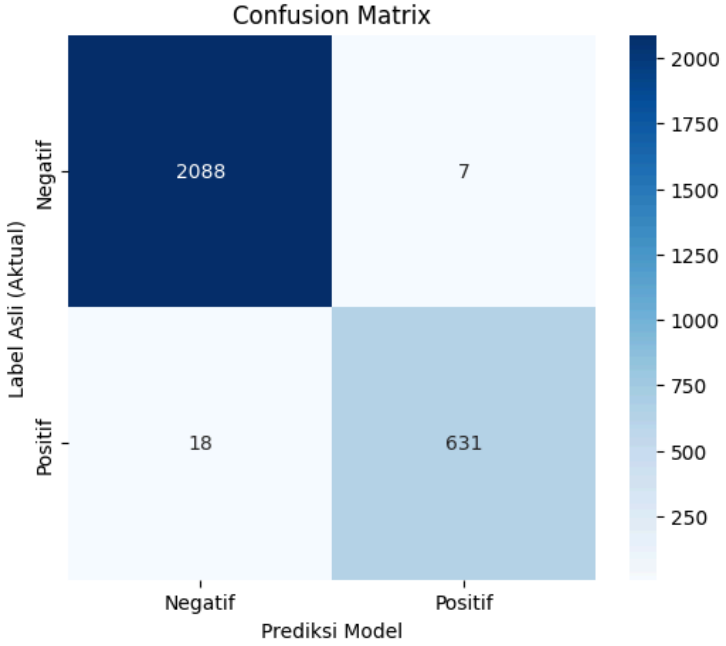
Kemudian kami juga melakukan uji pada dataset testing yang kami seimbangkan menggunakan WeightedRandomSampler dengan best parameter dari hasil hyperparameter tuning sebagai berikut:

```
Best Hyperparameters:
LR=2e-05, Batch Size=32, Epochs=5
Accuracy: 0.9901603498542274
F1 Score: 0.9901257944654389
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      2095
           1       0.99      0.97      0.98       649

    accuracy                           0.99      2744
   macro avg       0.99      0.98      0.99      2744
weighted avg       0.99      0.99      0.99      2744
```

Model dievaluasi dengan menggunakan best parameter diantaranya learning rate 2e-5, batch size berukuran 32 batch, dan 5 epochs. Akurasi model mencapai 0.99 menandakan hampir seluruh kelas dataset berhasil diprediksi secara tepat dari keseluruhan dataset, menunjukkan model dapat melakukan prediksi dengan baik. Hasil akurasi yang dihasilkan model dengan dataset yang sudah seimbang dengan WRS lebih tinggi dibandingkan dataset dengan model baseline yang tidak dilakukan penyeimbangan data dan parameter. Namun, terdapat penurunan precision untuk kelas hoax (1) dimana pada model baseline sebelumnya mencapai 1.0 namun turun menjadi 0.99 pada model dengan WRS, menunjukkan terdapat 1% data yang diprediksi hoax bukan merupakan kelas hoax. Sedangkan, terdapat peningkatan pada recall kelas hoax, dimana recall mencapai 0.97, yang sebelumnya pada model baseline hanya 0.96. Kenaikan recall menandakan model berhasil menambah jumlah data hoax yang berhasil diprediksi sebagai data hoax yang secara tepat merupakan data hoax dan hanya 3% data hoax yang lolos dimisklasifikasi sebagai data valid. F1-score memiliki nilai yang sama dengan base model, namun karena adanya peningkatan model dalam melakukan prediksi secara tepat untuk data hoax, menunjukkan dengan menghandle imbalance data dan menggunakan best parameter, performa model sedikit meningkat dari baseline model.




Dari sini dapat kita lihat dengan menghandle imbalance data, jumlah data hoax yang dimisklasifkasi sebagai data valid lebih sedikit dibandingkan sebelumnya, dimana awalnya terdapat 50 data yang diklasifikasikan sebagai data valid padahal sebenarnya data hoax, kini menurun tersisa hanya 18 data yang dimisklasifikasikan sebagai data valid.

Terakhir, kami melakukan evaluasi model pada dataset yang telah diseimbangkan dengan cara melakukan cutting dataset kelas mayoritas agar sesuai dengan jumlah kelas

minoritas. Dari evaluasi model dengan dataset yang sudah kami balance dari awal dengan menggunakan best parameter dari hasil hyperparameter tuning, kami memperoleh:

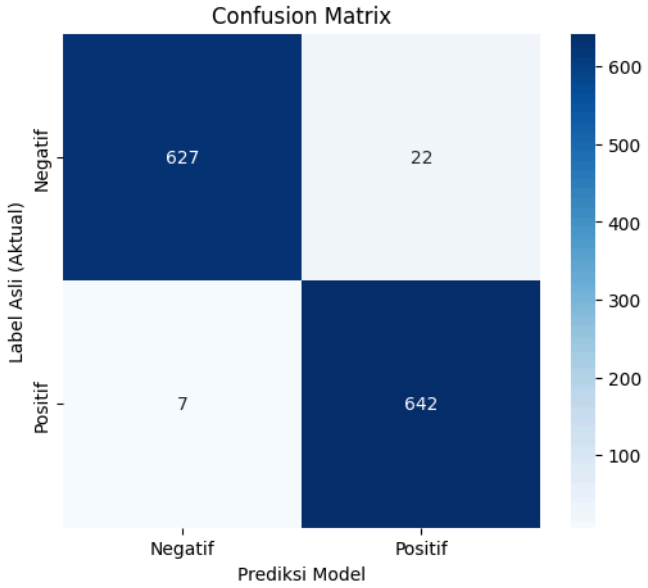
```
Mulai Evaluasi...
Accuracy : 0.9777
F1 Score : 0.9777

Detail Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.99      0.97      0.98       649
           1       0.97      0.99      0.98       649

    accuracy                           0.98      1298
   macro avg       0.98      0.98      0.98      1298
weighted avg       0.98      0.98      0.98      1298
```

Hasil akurasi yang dihasilkan model dengan dataset yang sudah balance sedari awal, mencapai 0.98 menandakan model memiliki performa yang masih berada dibawah model dengan handle imbalance WRS. Jika kita amati, terdapat penurunan precision pada kelas hoax (1) dimana sebanyak 3% data yang diprediksi hoax, sebenarnya bukan merupakan kelas hoax. Hal ini menunjukkan dengan melakukan training pada dataset yang sudah balance sedari awal dengan data uji yang balance, model sedikit mengalami kebingungan dalam melakukan klasifikasi data valid yang diklasifkasi sebagai hoax. Hal ini terlihat dari nilai recall dengan angka 0.97 mengindikasikan sebanyak 3% data valid salah diprediksi sebagai kelas hoax. Berdasarkan hasil metric evaluasi pada model dengan dataset yang sudah seimbangkan sedari awal, membuat model sedikit kebingungan dalam membedakan data valid yang diklasifkasikan sebagai hoax.



Pada confusion matrix untuk model dengan dataset yang dicut mengikuti jumlah kelas minoritas, terdapat 22 data valid yang dimisklasifkasikan sebagai data hoax, berbeda dengan model-model sebelumnya dimana lebih banyak data hoax yang diklasifkasikan sebagai data valid. Menunjukkan model mengalami penurunan performa apabila dataset kelasnya diseimbangkan sedari awal, dimana mengindikasikan model masih kurang memahami pola berita valid yang mengakibatkan adanya misklasifikasi data valid yang diklasifkasikan sebagai hoax.

## 10. Interpretation & Insights

Berdasarkan hasil prediksi terhadap artikel berita politik yang telah di scrapping dari Tribunnews, model mengklasifikasikan sebagian besar artikel sebagai berita valid, sementara sebagian lainnya diklasifikasikan sebagai hoax. Proporsi ini menunjukkan bahwa model tidak hanya memprediksi satu kelas secara dominan, melainkan mampu membedakan variasi konten yang ada dalam dataset. Tingkat confidence yang relatif tinggi pada sebagian besar prediksi mengindikasikan bahwa model memiliki keyakinan yang kuat terhadap keputusan klasifikasinya. Berdasarkan hasil prediksi tersebut, artikel yang diprediksi sebagai berita valid memiliki beberapa karakteristik umum seperti: mengandung bahasa formal dan informatif; menyebut tokoh politik atau partai secara eksplisit; dan menggunakan judul yang deskriptif dan tidak provokatif. Sedangkan untuk artikel berita yang diprediksi sebagai berita hoax juga memiliki beberapa karakteristiknya seperti menggunakan judul yang sensasional; minimnya rujukan terhadap sumber asli; dan narasi yang bersifat opini dibandingkan laporan faktual.

Sebagian besar artikel yang membahas pernyataan resmi pejabat negara, dinamika partai politik, atau kebijakan publik diprediksi secara konsisten sebagai berita non-hoax. Dari sini, kita mengetahui bahwa model mampu menangkap pola bahasa dan struktur khas berita politik yang faktual. Namun, terdapat juga beberapa artikel berita yang diprediksi sebagai hoax yang memiliki konteks politik yang sah, namun disampaikan dengan gaya bahasa yang ambigu atau spekulatif. Hal ini mengindikasikan bahwa model masih kesulitan dalam membedakan antara berita politik faktual dan opini politik yang dibungkus dalam format berita. Hal ini mencerminkan kompleksitas dari struktur bahasa yang digunakan dalam berita khususnya dalam hal politik.

Berdasarkan hasil interpretasi, terdapat beberapa insight penting yang bisa didapatkan. Model lebih andal dalam mengenali berita politik yang bersifat faktual dan struktural sehingga mampu memberikan hasil prediksi kelas berita yang lebih optimal. Selain itu, gaya bahasa dan konteks berita memiliki pengaruh yang besar terhadap hasil prediksi. Penggunaan diksi emosional, provokatif, dan sensasional cenderung mempengaruhi keputusan model dalam mengklasifikasikan berita. Artikel dengan konteks yang ambigu, seperti berita yang memadukan fakta dengan opini atau sindiran, menjadi tantangan utama bagi model dalam proses klasifikasi. Ambiguitas ini

menyebabkan batasan antara berita non-hoax dan berita hoax menjadi kurang tegas sehingga meningkatkan potensi kesalahan prediksi.

## 11. Final Output and Deliverables

Final output dari projek ini merupakan sebuah pipeline model untuk melakukan klasifikasi berita politik valid dan hoax berbahasa Indonesia. Pipeline model ini sudah mencakup proses pembersihan text untuk berita hoax, dan normalisasi data teks yang sebelumnya tidak terstruktur. Selain itu, pipeline juga sudah mencakup contextual embedding dengan indobert-base-p1 yang sudah melalui proses fine-tuning dengan data berita yang kami kumpulkan. Untuk penggunaan model, model IndoBert klasifikasi berita kami dapat digunakan langsung dengan melakukan load model yang sudah kami save sebelumnya. Seluruh parameter yang digunakan untuk pembentukan model baik untuk dataset imbalance, best parameter, dan untuk dataset yang sudah balance tersimpan dalam format *json* yang dapat ditemukan di dalam link repository github yang telah dilampirkan. Selain model, source code notebook, dataset, dokumentasi, pipeline untuk melakukan preprocessing, serta tools scrapping telah kami simpan ke dalam Github repository. Repository terbuka untuk publik sehingga seluruh pipeline baik model maupun text preprocessing dapat dijalankan dengan melakukan clone repository untuk pengembangan lebih lanjut.

## 12. Conclusion

Dengan kombinasi dataset yang berkualitas dan penggunaan model BERT yang memiliki kemampuan representasi bahasa yang baik, model prediksi ini mencapai performa akurasi yang optimal dengan angka mencapai 98.98% dan f1-score sebesar 98.97%. Hal ini didapatkan dari implementasi Weight Random Sampling sebagai upaya handle imbalance. Selain metode ini, kami juga melakukan cutting data dimana data berita non-hoax di potong agar sama banyaknya dengan data berita hoax. Dengan penerapan metode cutting data ini, didapatkan akurasi dan f1-score sebesar 97.77%.

Secara keseluruhan, proyek ini berhasil mengimplementasikan pendekatan *deep learning* berbasis IndoBERT untuk melakukan klasifikasi berita secara efektif. Melalui tahapan pengolahan data, pembentukan representasi teks, serta proses *fine-tuning* model, sistem yang dibangun mampu menangkap konteks bahasa Indonesia dengan baik dan menghasilkan performa yang memadai dalam membedakan berita valid dan hoaks. Hasil yang diperoleh menunjukkan bahwa pemanfaatan model *pre-trained* seperti IndoBERT merupakan solusi yang relevan dan kuat untuk permasalahan klasifikasi teks, khususnya pada domain berita politik Indonesia, sekaligus menegaskan pentingnya integrasi antara kualitas data, metode pemodelan, dan perancangan eksperimen yang sistematis dalam menghasilkan model prediksi yang andal. Namun, walaupun model klasifikasi

menunjukkan performa yang sangat baik, kebenaran fakta suatu berita tetap perlu ditelusuri kebenaran dan keabsahannya secara langsung, karena model klasifikasi ini hanya menentukan kelas berita berdasarkan pola linguistiknya saja tanpa mengecek langsung fakta atau sumber beritanya.

## 13. Future Improvements

Meskipun model klasifikasi IndoBERT-based kami telah mencapai performa yang cukup baik, peningkatan selanjutnya akan berfokuskan untuk membangun model yang lebih robust dan dapat digeneralisasikan pada skenario berita secara nyata. Karena adanya keterbatasan dataset berita hoax, kami menyarankan untuk melakukan analisis lanjut terhadap data hoax yang dimisklasifikasi sebagai valid dan sebaliknya. Analisis ini dibutuhkan guna memahami pola kalimat yang membingungkan model dalam melakukan pengklasifikasian terhadap berita valid dan hoax. Peningkatan selanjutnya dapat melakukan pengujian model terhadap sumber berita dengan topik yang berbeda, untuk mengecek apakah model overfit terhadap gaya penulisan tertentu. Selain itu, penelitian dengan membandingkan model IndoBERT dengan model lain akan sangat membantu dalam menemukan model yang paling cocok untuk melakukan klasifikasi berita valid dan hoax.

## 14. Reflection

**Novellina Edyawati**

Dalam mengerjakan final project text mining ini, saya belajar berbagai pengetahuan baru, khususnya karena kami mengangkat topik klasifikasi berita valid dan berita hoax. Mulai dari menentukan topik, mencari jurnal terkait dengan topik dan arsitektur model klasifikasi yang dibangun, dan berbagai tahapan proses pembentukan model, mulai dari EDA, preprocessing, tokenizing, building model, melakukan prediksi data, sampai evaluasi model, dan menyusun laporan. Melalui berbagai tahapan pengerjaan model diatas saya belajar untuk mengatur waktu bagaimana menyusun model sambil melakukan berbagai pekerjaan lain, selain itu lewat project ini mengasah skill analisis saya baik ketika melakukan exploratory data analisis dan hasil evaluasi model. Dimana dari exploratory data analisis dan evaluasi model ini saya menemukan bahwa berita hoax memiliki pola tersendiri dimana gaya penulisan bahasanya biasanya menggunakan bahasa sensasional yang dapat memicu kontroversi seperti kata korban, penjelasan beredar, dsbg. Dimana kata-kata tersebut tidak memiliki sumber resmi, sehingga berita hoax biasanya akan membuat kalimat klarifikasi menggunakan kata seperti berdasarkan hasil penelusuran, klarifikasi, dsbgnya. Selain itu untuk membangun model indoBert sendiri kami perlu membaca dokumentasi dari indolem untuk memilih model dan tokenizer LLM yang cocok dengan dataset kami. Dalam menjalankan proyek ini,

tantangan yang paling menonjol bagi kami adalah waktu komputasi ketika melakukan training dan hyperparameter tuning model karena hal ini memerlukan waktu yang lama serta memori yang sangat banyak. Hal ini menjadi tantangan besar karena dengan waktu komputasi yang cukup lama, berulang kali model kami stuck dan perlu mengulang training model, sehingga untuk proses evaluasi model jadi terhambat. Bagian building model ini yang paling memakan waktu karena kami membutuhkan waktu komputasi yang cukup lama serta kami juga perlu melakukan pengoptimalan model dengan melakukan balancing dataset baik dengan WRS dan cutting dataset, serta mencari hyperparameter yang cocok supaya performa model kami dapat optimal. Lewat proyek ini saya belajar lebih dalam mengenai teknis penerapan NLP lewat deeplearning model IndoBERT untuk mengolah dataset berbahasa Indonesia serta penerapan text representation yang sangat penting bagi model deep learning agar bisa mendapatkan makna kalimat sehingga bisa melakukan klasifikasi berita hoax dan berita valid dengan baik. Selain itu, saya juga belajar lewat kolaborasi dalam kelompok kami, banyak aspek yang mungkin saya lewatkan tapi karena berkat diskusi dan bantuan dari partner saya, saya dapat menemukan aspek ataupun insight yang terlewat dan mendapat masukan untuk membangun model kami menjadi lebih baik.

**Sherly Oktavia Willisa**

Selama pengerjaan, saya memperoleh pengalaman yang sangat bermakna dan memperluas pemahaman saya secara signifikan, khususnya dalam membangun model IndoBERT untuk melakukan prediksi berita hoax dan berita valid. Melalui proyek ini, saya mengetahui bahwa membangun sebuah model harus melalui rangkaian tahapan yang saling berkaitan dan membutuhkan ketelitian sejak awal. Mulai dari eksplorasi jurnal serupa untuk mengetahui model apa yang digunakan oleh peneliti lain, kemudian mencari dataset yang sesuai agar memberikan hasil prediksi yang maksimal dan optimal, hingga mempelajari cara mengolah data hingga akhirnya menjadi sebuah model prediksi merupakan suatu ilmu baru yang saya peroleh melalui proyek ini. Tantangan yang paling menonjol selama pengerjaan proyek ini terletak pada aspek komputasi dimana waktu komputasi yang diperlukan memakan waktu yang cukup lama dan sumber daya yang cukup besar. Hal ini menjadi hambatan dalam pengerjaan proyek kelompok ini karena spesifikasi device kami yang terbatas. Dari sini, saya menyadari bahwa proyek berbasis Deep Learning idealnya direncanakan dan dimulai sejak jauh hari agar alokasi waktu dan sumber daya dapat dimanfaatkan secara lebih optimal. Bagian inilah yang menjadi tahapan paling memakan waktu sekaligus paling menantang dalam keseluruhan proyek. Dari seluruh proses tersebut, saya merasa bahwa kami mempelajari aspek teknis, seperti penerapan text representation dan permodelan dengan IndoBERT, juga bagaimana kolaborasi tim menjadi aspek penting dalam kerja kelompok. Melalui kesempatan ini, saya mampu melatih kemampuan manajemen proyek, terutama mengatur waktu,

menyusun prioritas, dan menjaga koordinasi agar seluruh tahapan dapat berjalan secara terstruktur.

## 15. References


Rizqullah, M. R., Radhinansyah H. G., Steve I. H., "Indonesian Fact and Hoax Political News," Kaggle Dataset, 2021. [Online]. Available: https://www.kaggle.com/datasets/linkgish/indonesian-fact-and-hoax-political-news

Yopita Desriana Butar. (2024). Analisis Penyebaran Hoax Di Media Sosial Dan Dampaknya Terhadap Masyarakat. *Jurnal Pendidikan, Bahasa Dan Budaya*, *3*(2), 252–258. https://doi.org/10.55606/jpbb.v3i2.3201

Koto, F., Rahimi, A., Lau, J., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP* (pp. 757–770). https://aclanthology.org/2020.coling-main.66.pdf

Putra, F., & Patra, H. (2023). Analisis Hoax pada Pemilu: Tinjauan dari Perspektif Pendidikan Politik. *Naradidik: Journal of Education and Pedagogy*, *2*(1), 95–102. https://doi.org/10.24036/nara.v2i1.119

Marwan, M., Jurusan, A., Komunikasi, I., Fakultas, I., & Komunikasi. (n.d.). *ANALISIS PENYEBARAN BERITA HOAX DI INDONESIA*. http://download.garuda.kemdikbud.go.id/article.php?article=916041&val=9538&title=Sistem%20Informasi%20Penilaian%20Pemberitaan%20Hoax%20dengan%20Metode%20Perbandingan%20Dan%20Algoritma%20AHP

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). https://doi.org/10.33299/jpkop.23.1.1805

Asep Ripa'i, Santoso, F., & Farihin Lazim. (2024). Deteksi Berita Hoax dengan Perbandingan Website Menggunakan Pendekatan Deep Learning Algoritma BERT. *Jurnal Teknologi Terapan G-Tech*, *8*(3), 1749–1758. https://doi.org/10.33379/gtech.v8i3.4541

Prasetya, F., & Ferdiansyah, F. (2022). Analisis Data Mining Klasifikasi Berita Hoax COVID 19 Menggunakan Algoritma Naive Bayes. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(1), 132. https://doi.org/10.30865/json.v4i1.4852

Handayani, Nur and Amir, Johar and Juanda, Juanda (2021) *KASUS HOAKS PANDEMI COVID-19: SUATU TINJAUAN LINGUISTIK FORENSIK*. Fon : Jurnal Pendidikan Bahasa dan Sastra Indonesia, 17 (2). pp. 169-177. ISSN p-ISSN 2086-0609 e-ISSN 2614-7718

P. E. G. Yasa, "Ujaran Kebencian dan Hoax Perspektif Karya Sastra Klasik: Kajian Linguistik Forensik," in *Prosiding Seminar Nasional Bahasa, Sastra, & Budaya*, Vol. 2, 2023, pp. [if available], Univ. Udayana, Bali, Dec. 27, 2023. [Online]. Available: https://ejournal1.unud.ac.id/index.php/snbsb/article/view/774