



## TURNAMEN SAINS DATA NASIONAL 2023

# AI Fake News Detection

Web Apps about classification of fake or real news

Nama penyusun

1. Muhammad **Iqbal**
2. **Lutfi** Lingga Saputra
3. **Clara** Natalie Surjadajadi
4. Sekar Ayu **Larasati**



# DISPONSORI OLEH

LSP SAINS DATA DAN  
KECERDASAN BUATAN  
INDONESIA





# DIDUKUNG OLEH



**panrb**

KEMENTERIAN  
PENDAYAGUNAAN APARATUR NEGARA  
DAN REFORMASI BIROKRASI



**djp**



**Satu Data  
JAKARTA**

 **jakarta  
smart city**



**JABAR  
DIGITAL  
SERVICE**



**OPEN  
DATA  
JABAR**

Asosiasi  
Data Sains dan AI  
**Indonesia**



**DATA  
SCIENCE  
INDONESIA**

**ABI**  
ASOSIASI BIG DATA INDONESIA

 **BigData**



# DIDUKUNG OLEH



UNIVERSITAS  
ISLAM  
INDONESIA



UNIVERSITAS  
BUDI LUHUR



UNIVERSITAS  
KUNINGAN



TURNAMEN SAINS DATA NASIONAL 2023



# DAFTAR ISI

- 1. Pendahuluan\_\_\_\_\_ Slide 7**
- 2. Latar Belakang\_\_\_\_\_ Slide 8**
- 3. Analisa Kebutuhan Data\_\_\_\_\_ Slide 11**
- 4. Proses Flow\_\_\_\_\_ Slide 12**
- 5. Screen Capture\_\_\_\_\_ Slide 20**
- 6. Kesimpulan dan Saran\_\_\_\_\_ Slide 28**
- 7. Daftar Pustaka\_\_\_\_\_ Slide 30**

# INFORMASI TIM

Bayez adalah tim yang terdiri dari 4 orang, yang bekerja di suatu project bersama. Project yang pernah dikerjakan yakni berkaitan dengan data science seperti data visualization, supervised dan unsupervised learning.



Nama : Muhammad **Iqbal**  
Jabatan : Data Scientist  
Medsos : iqbalaws (IG)  
Hobi : Listening Podcast



Nama : **Clara Natalie Surjadajadi**  
Jabatan : Data Scientist  
Medsos : claranatalies (IG)  
Hobi : badminton



Nama : **Lutfi Lingga Saputra**  
Jabatan : Data Scientist  
Medsos : ltflngg (IG)  
Hobi : Nonton Film



Nama : **Sekar Ayu Larasati**  
Jabatan : Data Scientist  
Medsos : sa\_larasati (IG)  
Hobi : Membaca Buku

# 1. Pendahuluan

**Hoaks** adalah suatu penyampaian informasi yang diberitakan namun tidak sesuai dengan senyatanya, beritanya sudah diramu sedemikian rupa dengan menambahkan atau mengurangi isi dari berita. Pemicu terjadinya informasi hoaks dipicu dua motif yaitu ekonomi dan politik (Himslaw Article, 2020).



Berdasarkan data World Press Freedom Index per 2023, Indonesia berada di posisi 108 dari 180 negara (RSF, 2023) di mana independensi media masih belum maksimal untuk menyajikan data yang dapat dipercaya oleh publik. Hal ini menjadi salah satu proxy yang digunakan untuk mengukur tingkat hoaks yang beredar di Masyarakat.

Beberapa kanal berita: media jurnalistik, media cetak, radio, televisi, hingga website. Keseluruhan kanal berita tersebut rentan terhadap hoax. (GDI, 2022). Bagi sebagian orang, agak sulit untuk membedakan mana berita yang benar dan mana berita yang bohong (hoaks). (Amilin, 2019)

Hoaks dapat menyebabkan perseteruan hingga dapat menimbulkan konflik dalam kehidupan masyarakat. Hal tersebut tentunya sangat rawan dan dapat menimbulkan bentrokan serta bisa mengganggu stabilitas kehidupan dalam masyarakat juga bisa mengancam keutuhan negara dan kebinekaan (Nuraisyah, 2017).

## 2. Latar Belakang

Berdasarkan Kominfo (2023) dari bulan Agustus 2018 sampai bulan Mei 2023, sebanyak 11.642 konten hoaks telah diidentifikasi. Sejumlah 1.373 diantaranya masuk ke dalam kategori politik.



Sumber : Kominfo, 2023

Pemanfaatan media sosial atau kanal berita lainnya guna kepentingan politik banyak disalahgunakan oleh sebagian orang tertentu untuk merebut perhatian dan simpati masyarakat. Portal berita yang seharusnya digunakan untuk melakukan literasi agar masyarakat Indonesia paham tentang politik, justru oleh sebagian orang digunakan sebagai media propaganda dan provokasi (Amilin, 2019).

Berdasarkan Amilin (2019) Hoaks politik perlu dimitigasi dengan cara yang baik, benar, dan tepat. Salah satunya adalah negara perlu memberikan solusi cerdas menghadapi perkembangan teknologi informasi dan komunikasi.

## 2. Latar Belakang (2)



Contoh Headline atau judul konten HOAX

Sumber : TurnBackHoax,  
2023

Konten **Hoax** kebanyakan memiliki headline yang sangat provokatif dan juga sekaligus menarik perhatian pembaca, sehingga banyak sekali pembaca yang terprovokasi atas konten Hoax tersebut, hal ini bisa dimanfaatkan oleh kreator Hoax untuk menghasut atau bahkan melakukan Phising dengan cara menyebar judul konten yang disertai link lengkap.

Berdasarkan latar belakang tersebut, disini kami memilih menggunakan pendekatan Artificial Intelligence untuk mendeteksi mana konten Hoax atau Asli berdasarkan **headline** atau **judul** konten tersebut. Konten yang dimaksud disini bisa berasal dari konten video, foto maupun, tulisan di media sosial maupun headline dari portal berita.

## 2. Latar Belakang (3)

### BAGAIMANA CARA MENDETEKSI HOAX?

Rumusan Masalah :

1. Bagaimana cara mendeteksi Hoax?
2. Bagaimana model AI bisa untuk digunakan dalam *Hoax Detection*?
3. Bagaimana penerapan model tersebut diintegrasikan ke dalam Web Apps dalam *Hoax Detection*?

### APAKAH WEB APPS MERUPAKAN SOLUSI?



Alasan Menggunakan Web-Apps :

1. AI model bisa membantu user untuk menentukan mana berita Hoax atau tidak
2. Aplikasi Web Apps dapat diintegrasikan dengan mesin dan model AI pendekripsi Hoax yang kami buat
3. Aplikasi web Apps dapat diakses melalui browser, yang bisa di hosting secara publik dan diakses oleh semua orang.

### 3. Analisa Kebutuhan Data

	Title	label
0	Partai Buruh Bantah Terlibat Dugaan Manipulasi...	0
1	KPU Luncurkan Sipol untuk Pendaftaran Parpol d...	0
2	Mahfud Vs Rizal Ramli, Twitwar Caper Minim Sub...	0
3	Luhut Tolak Tawaran Jadi Cawapres Pendamping A...	0
4	DPR Ubah Tatib terkait Pansus RUU IKN, Pengama...	0
...	...	...
21995	Orang-orang China di Terminal 1C Cengkareng Ma...	1
21996	Pengacara Hotman Paris Hutapea Jadi Kuasa Huku...	1
21997	Aksi Kita Indonesia Tandingan dari Aksi Bela I...	1
21998	Foto pejabat keuangan dibawah palu arit	1
21999	Kaesang Bapak Saya dengan Kesederhanaan Bisa Nip...	1

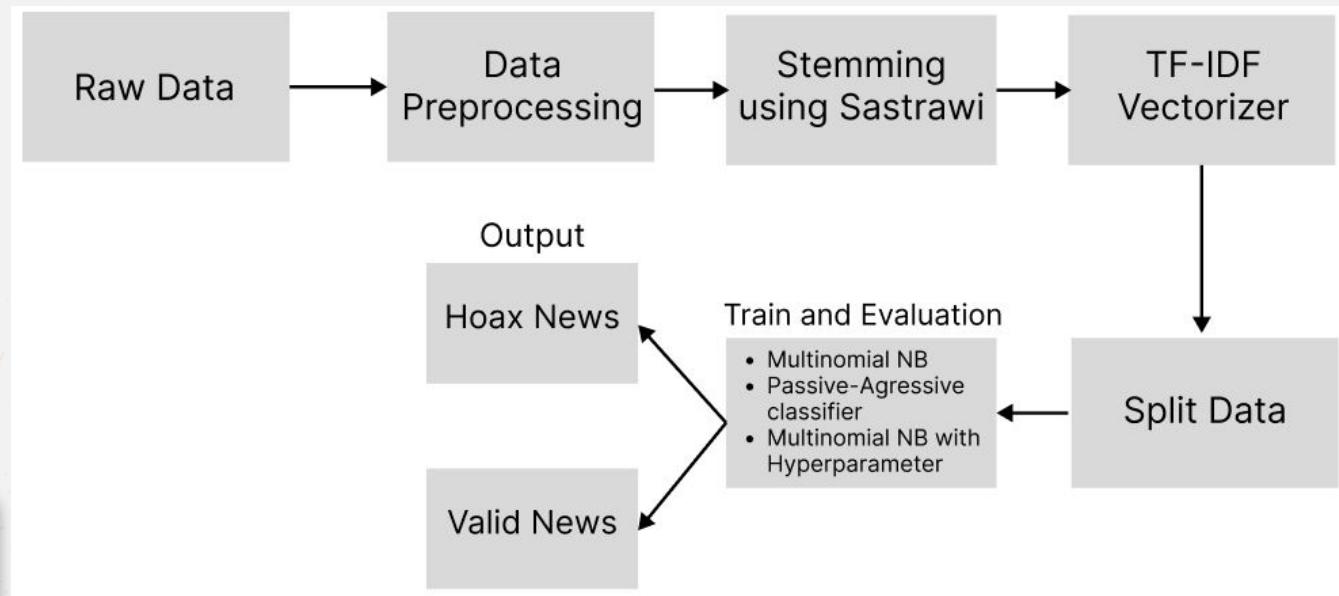
[22000 rows x 2 columns]

Dataset yang kami gunakan adalah dataset berita politik dan non politik Indonesia yang bersifat campuran yang terdiri atas banyak sumber, melibatkan berita hoax dan berita yang benar. **Dataset ini terstruktur dalam format yang terdiri dari 22.000 baris, di mana setiap baris merepresentasikan satu berita politik Indonesia.**

<https://www.kaggle.com/datasets/linkgish/indonesian-fact-and-hoax-political-news>

# 4. Proses Flow (1)

## Illustration of Machine Learning (ML)'s Architecture Model



# 4. Proses Flow (2)

## Penjelasan Proses Flow ML:

### I. Raw Data:

Dataset bersumber dari berita politik dan non politik di Indonesia yang terdiri atas banyak sumber, yang terdiri dari berita hoax dan berita yang valid. Dataset ini terstruktur dalam format yang terdiri dari 22.000 baris, di mana setiap baris merepresentasikan satu berita politik Indonesia.

### II. Data Preprocessing:

Pada bagian ini, dilakukan berbagai perlakuan terhadap data yang digunakan pada seluruh dataset seperti: penghapusan data yang berduplikat, penghapusan kata yang tidak diperlukan melalui keywords, pemberian label pada dataset (0=valid, 1=hoax), penghapusan kata penghubung di seluruh dataset.

### III. Stemming using Sastrawi:

Pada bagian ini, dilakukan penginstallan Sastrawi yang dilakukan untuk pemberian makna pada kata-kata berbahasa Indonesia yang ada pada dataset.

# 4. Proses Flow (3)

## Penjelasan Proses Flow ML:

### IV. TF IDF Vectorizer:

Vectorizer ini dipakai untuk melihat seberapa penting nilai sebuah kata dalam suatu corpus atau kalimat text berdasarkan kuantitas kemunculan dan pengaruhnya terhadap klasifikasi.

### V. Split Data:

Membagi data menjadi 2 bagian, yaitu training 80% untuk melatih model dan test 20% untuk uji performa model. Training juga dibagi menjadi 2 bagian lagi yaitu 60% dan 20% untuk memastikan semua bagian dari data terlatih dengan porsi yang sesuai (cross validation).

### VI. Train and Evaluation:

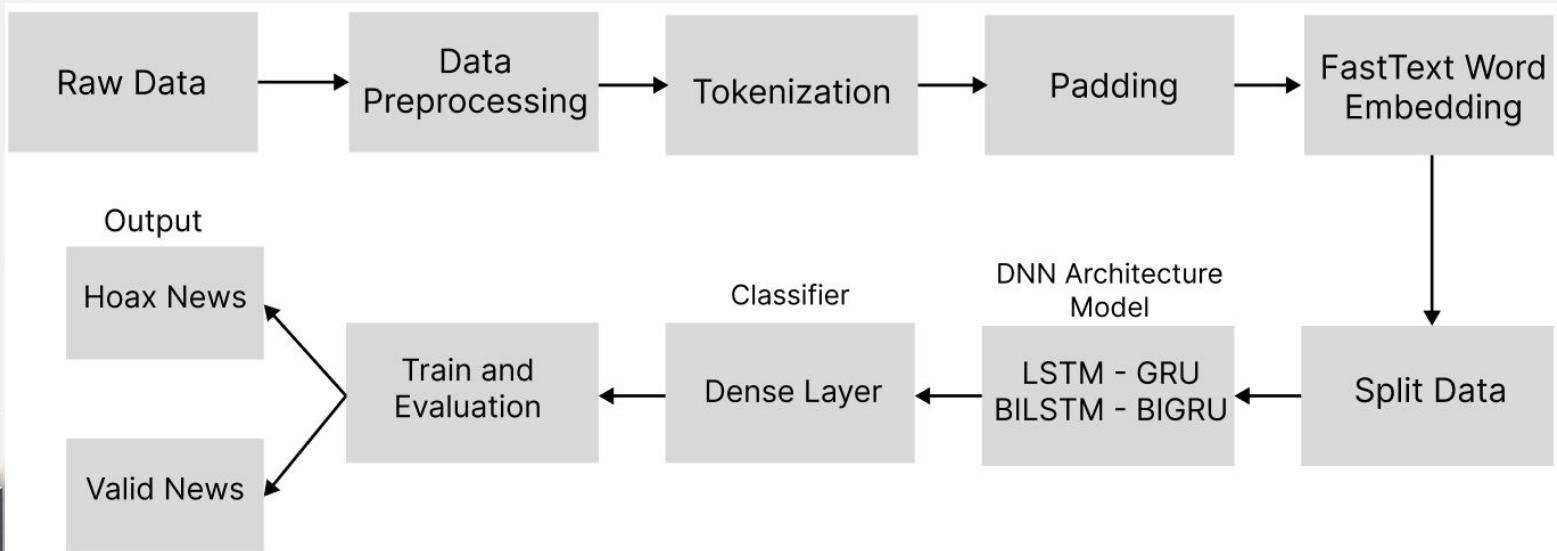
Mencoba membentuk model dengan 3 algoritma, yaitu **Multinomial NB**, **Passive Aggressive** dan **multinomial NB hyperparameter**. Lalu model masing-masing di evaluasi dengan metrics seperti **Accuracy**, **F1**, **Recall** dan **Precision**.

### VII. Output:

Model dengan performa terbaik akan dijadikan backend atau mesin dalam WebApps.

## 4. Proses Flow (4)

Illustration of Deep Neural Network (DNN)'s Architecture Model



# 4. Proses Flow (5)

## Penjelasan Proses Flow DNN:

### I. Raw Data:

Dataset bersumber dari berita politik dan non politik di Indonesia yang yang terdiri atas banyak sumber, yang terdiri dari berita hoax dan berita yang valid. Dataset ini terstruktur dalam format yang terdiri dari 22.000 baris, di mana setiap baris merepresentasikan satu berita politik Indonesia.

### II. Data Preprocessing:

Menggantikan non-alfabet dengan spasi, merubah ke huruf kecil dan Kata-kata dipecah menjadi list, lalu menghapus stopwords dalam Bahasa Indonesia. Hasilnya disimpan dalam "corpus".

### III. Tokenization:

Mengonversi kata dalam "corpus" menjadi token numerik. fit\_on\_texts mempelajari vocabulari dari "corpus", dan texts\_to\_sequences mengubah setiap kata dalam "corpus" menjadi token numerik yang sesuai.

# 4. Proses Flow (6)

## Penjelasan Proses Flow DNN:

### IV. Padding:

Memastikan bahwa setiap sequence input memiliki panjang yang sama dengan cara menstandardisasi panjang sequence menjadi 229, menambahkan 0 jika pendek dan memotong jika terlalu panjang.

### V. Fast-text Word Embedding:

Membuat matriks embedding, di mana setiap barisnya sesuai dengan sebuah kata dari indeks kata tokenizer, dan setiap nilai baris adalah vektor FastText untuk kata tersebut.

### VI. Split Data:

Membagi data menjadi 2 bagian, yaitu training 80% untuk melatih model dan test 20% untuk uji performa model. Training juga dibagi menjadi 2 bagian lagi yaitu 60% dan 20% untuk memastikan semua bagian dari data terlatih dengan porsi yang sesuai (cross validation).

### VII. DNN's Architecture Model:

Membuat model sequential menggunakan Keras dengan 4 jenis layer RNN yang berbeda (LSTM, GRU, BI-LSTM, dan BI-GRU)

### VIII. Dense Layer:

Menggunakan fungsi aktivasi ReLU untuk mengekstrak fitur non-linear dan memahami pola yang kompleks dalam serta aktivasi sigmoid untuk menghasilkan output antara 0 dan 1, yang bisa diinterpretasikan sebagai probabilitas. Penggunaan dropout dan regularisasi L2 membantu mencegah overfitting

### IX. Train and Evaluation:

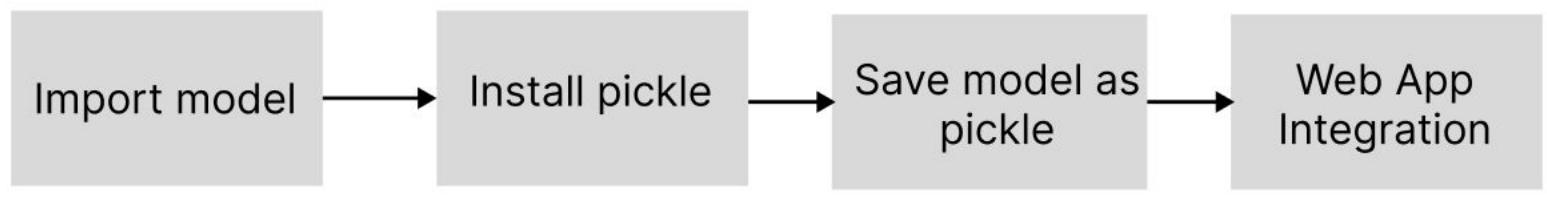
Model Sequential dengan layer RNN yang berbeda dievaluasi dengan metrics seperti Accuracy, F1, Recall dan Precision.

### X. Output:

Model Sequential dengan layer LSTM memiliki performa terbaik untuk diintegrasikan dalam WebApps.

## 4. Proses Flow (7)

### Illustration of Web-apps Model Integration



# 4. Proses Flow (8)

## Penjelasan Proses Web Apps Integration:

### I. Import model:

Melakukan import model yang akan dipakai untuk menjadi mesin pada webApps.

### II. Install pickle:

Install pickle pada jupyter untuk nantinya menjadikan model sebagai file ber format .pkl atau .h5

### III. Save model as pickle:

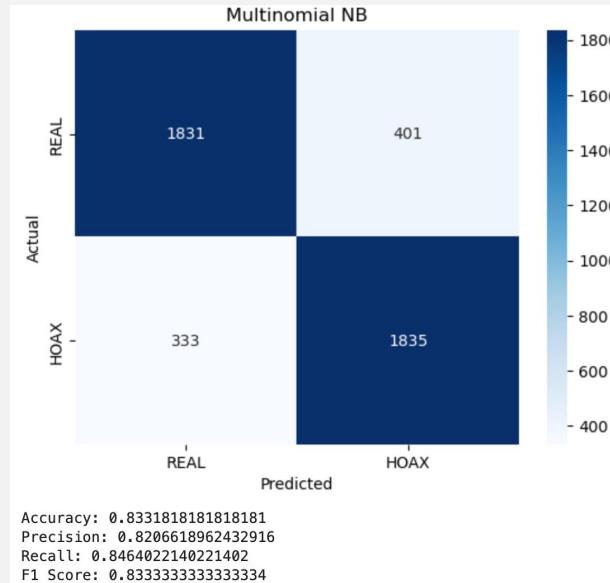
Save model dan vectorizer etc ke dalam format .pkl atau .h5

### IV. Web App Integration:

Buka Spyder, lalu integrasikan model ke dalam script format .py yang akan dipakai untuk menganalisis hasil input berita. Jangan lupa siapkan file .html dan .css untuk mengatur tampilan WebApps. Setelah itu, script berisi model di “Run” dan WebApps bisa ditampilkan di browser.

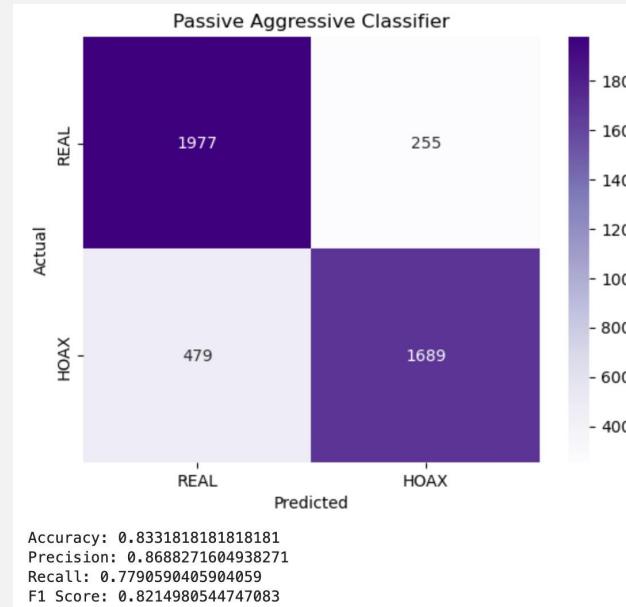
# 5. Screen Capture (1)

Hasil Run ML Multinomial Naive Bayes:



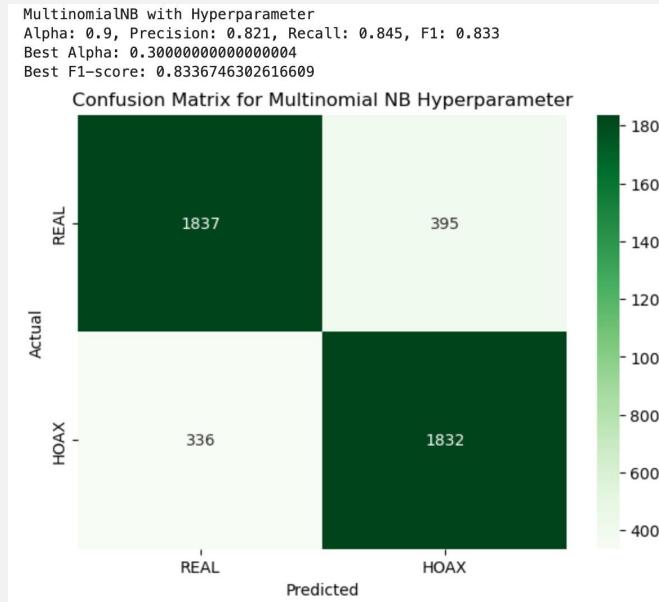
# 5. Screen Capture (2)

Hasil Run ML Passive-Aggressive Classifier:



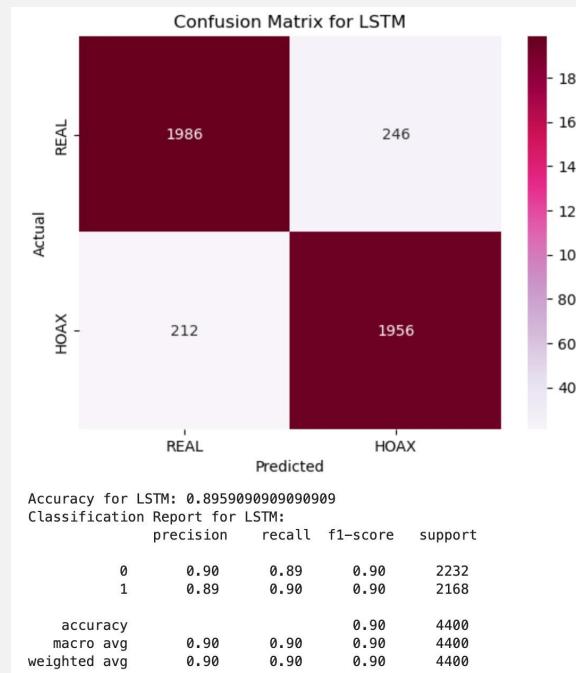
# 5. Screen Capture (3)

## Hasil Run ML Multinomial Naive Bayes with Hyperparameter:



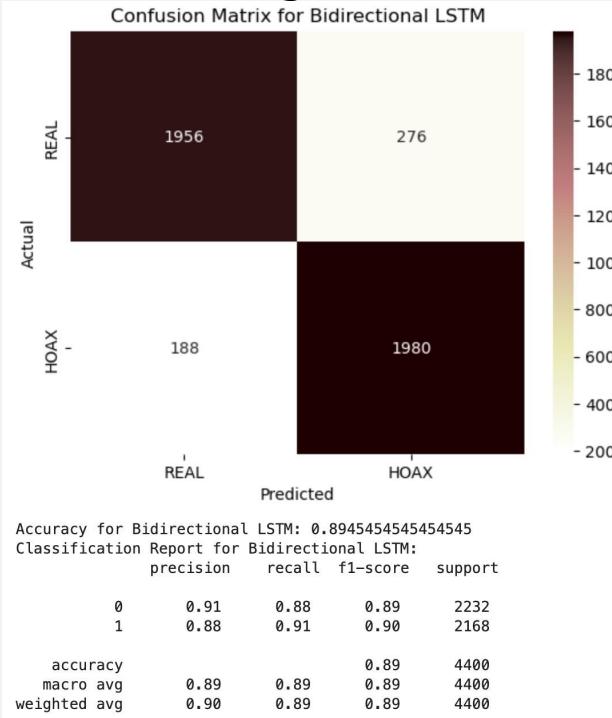
# 5. Screen Capture (4)

## Hasil Run DNN Long Short-Term Memory (LSTM):



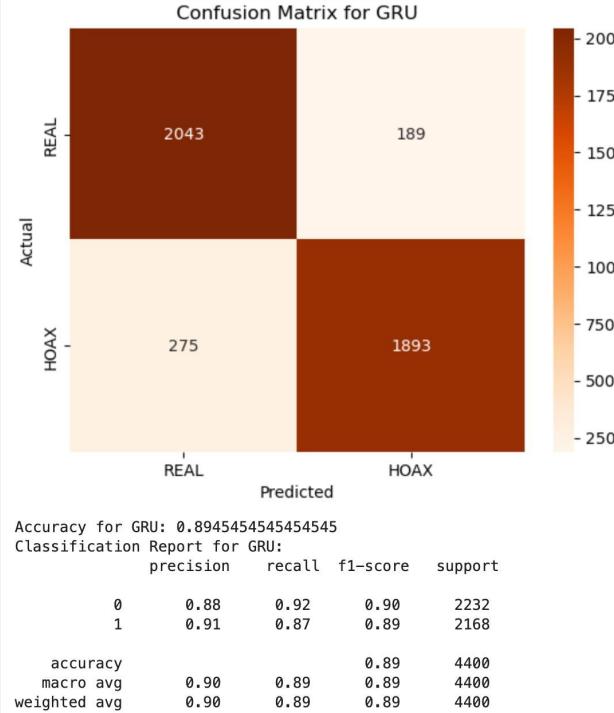
# 5. Screen Capture (5)

## Hasil Run DNN Bidirectional Long Short-Term Memory (Bi LSTM):



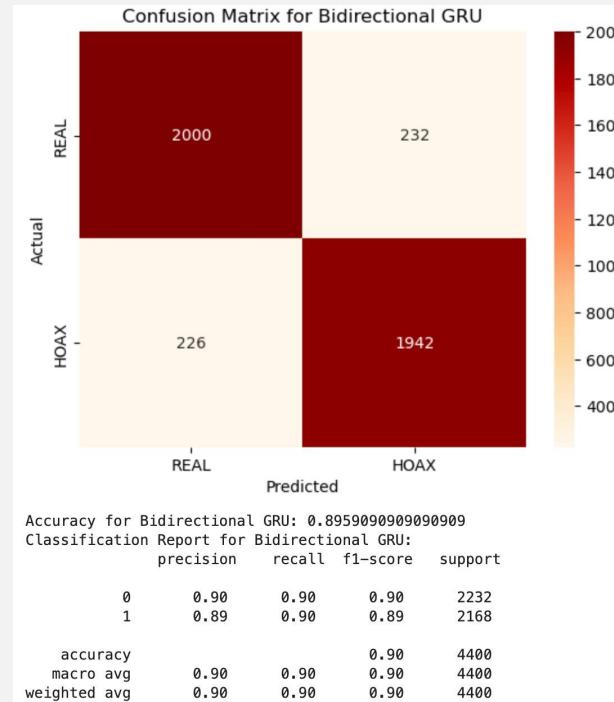
# 5. Screen Capture (6)

## Hasil Run DNN Gated Recurrent Unit (GRU):



# 5. Screen Capture (7)

## Hasil Run DNN Bidirectional Gated Recurrent Unit (Bi-GRU):



# 5. Screen Capture (8)

Tampilan Web Apps:



Fake News  
Detector

PKS dan Gerindra Purwakarta Bantah  
Pasang Spanduk yang Dinilai Provokatif

VALID CHECKING

## 6. Kesimpulan dan Saran (1)

1. Berdasarkan berbagai sumber jurnal ilmiah internasional, untuk mendeteksi Hoax lebih baik menggunakan *deep neural network* (DNN) sebagai toolsnya. Pada tulisan kami ini, didapatkan bukti bahwa penggunaan DNN memperoleh hasil yang lebih baik dibandingkan jika dibandingkan dengan proses yang menggunakan *Machine Learning* (ML);
2. Dari keseluruhan model dalam DNN, kami memperoleh hasil bahwa penggunaan *long short-term memory* (LSTM) merupakan model terbaik DNN untuk mendeteksi Hoax, jika dibandingkan model DNN lainnya seperti *bidirectional long short-term memory* (BiLSTM), *gated recurrent unit* (GRU), maupun *bidirectional long short-term memory* (BiGRU);
3. Penggunaan WebApps pada *hoax detection* dapat membantu masyarakat awam untuk lebih mudah mengetahui validitas suatu berita sehingga dapat memilah berita mana yang perlu dijadikan sumber acuan.

## 6. Kesimpulan dan Saran (2)

1. Setelah menyelesaikan tulisan ini, saran yang dapat kami berikan kepada masyarakat umum yang ingin melakukan penulisan karya ilmiah/penelitian lainnya yakni penggunaan dataset dapat diperluas tidak hanya yang berbahasa Indonesia saja, namun juga perlu memasukkan berita tentang isu-isu di Indonesia yang berbahasa Inggris, mengingat pencipta hoax tidak menutup kemungkinan hanya berasal dari sumber lokal, namun bisa juga dari media internasional;
2. Dapat dipertimbangkan juga untuk mengolah dataset yang bersumber khusus dari *social media*, seperti Twitter API atau Facebook API, mengingat peredaran hoax juga cukup masif di *social media* tersebut;
3. Selain itu, penulis selanjutnya juga dapat mencoba model terbaik lainnya menggunakan model lainnya pada *machine learning*, seperti *support vector machine* (SVM) ataupun *deep neural networks*, seperti *1-dimensional convolutional neural network* (1D-CNN).



# 7. Daftar Pustaka

Jurnal:

Amilin, A. "Pengaruh Hoaks Politik dalam Era Post-Truth terhadap Ketahanan Nasional dan Dampaknya pada Kelangsungan Pembangunan Nasional." *Jurnal Lemhannas RI* 7, no. 3 (2019): 5-11.

Nayoga, B. P., Adipradana, R., Suryadi, R., Soehartono, D. (2021). *Hoax Analyzer for Indonesian News Using Deep Learning Models*. *Procedia Computer Science* 179 (2021), 704-71.

Siddiq, N. (2017). Penegakan Hukum Pidana Dalam Penanggulangan Berita Palsu (Hoax) Menurut Undang-Undang No.11 Tahun 2008 Yang Telah Dirubah Menjadi Undang-Undang No.19 Tahun 2016 Tentang Informasi Dan Transaksi Lex Et Societatis Vol. V, No. 10, h. 26-32.

Situs:

Himslaw Article, Bahaya Menyebarluaskan Berita Hoaks, diakses Tgl 2 Februari 2020

<https://rsf.org/en/index>

<https://www.disinformationindex.org/country-studies/2022-11-02-disinformation-risk-assessment-the-online-news-market-in-indonesia/>

<https://turnbackhoax.id/>



**TERIMA KASIH**  
SEKIAN