# Klasifikasi Berita Bohong Berbasis Bahasa Indonesia Menggunakan Metode IndoBERT Pretrained Model


**Anggota:**

Pratama Azmi Atmajaya

Bagja 9102 Kurniawan

Fendi Irfan Amorokhman

**Asal Universitas:**

Telkom University

# DAFTAR ISI

# DAFTAR GAMBAR



## DAFTAR TABEL

# 1. Latar Belakang

Dewasa ini, kebutuhan informasi sudah menjadi kebutuhan umum masyarakat Indonesia, termasuk masyarakat yang tinggal di desa yang jauh dari kota. Perkembangan informasi sudah sudah merambat ke seluruh lapisan masyarakat dari atas, menengah, sampai kebawah pun sudah bisa menikmati informasi-informasi yang sangat luas. Hal ini dibuktikan dari riset yang dilakukan bulan januari 2020 lalu, yaitu 64% penduduk Indonesia sudah pakai internet.  Bukti ini menunjukkan bahwasanya masyarakat secara garis besar menerima perkembangan teknologi dan berusaha beradaptasi dengan kemajuan teknologi. Sayangnya, kebebasan dalam menyebarkan sebuah informasi dapat berdampak negatif, salah satunya ialah *hoax*. Berdasarkan hasil riset pada tahun 2020 ini, terjadi peningkatan *hoax* yang signifikan. Dikutip dari KOMINFO "Sejak 23 Januari 2020 hingga 15 Juni 2020 terdapat setidaknya 850 *hoax* yang beredar baik melalui media sosial maupun aplikasi pesan instan".[1] Dikutip dari KOMINFO "Data Kemenkominfo menyebutkan bahwa ada sekitar 800.000 situs di Indonesia yang telah terindikasi sebagai penyebar informasi palsu".[2] *Hoax* tidak dapat dipandang sebelah mata, nyatanya *hoax* dapat membuat chaos. Sebagai contoh *hoax* yang terjadi baru-baru ini, yaitu RUU Cipta kerja.

Menurut KBBI *hoax* bermakna berita bohong, berita tidak bersumber. (Kemendikbud, 2019). *Hoax* adalah informasi sesat dan berbahaya karena menyesatkan persepsi manusia dengan menyampaikan informasi palsu sebagai kebenaran. (Afriza & Adi Santoso, 2018) Secara garis besar *hoax* adalah berita yang menyesatkan karena tidak mempunyai sumber yang dapat dipertanggungjawabkan dan bukti yang jelas. Berita *hoax* sengaja diciptakan oleh segelintir orang untuk memperoleh keuntungan pribadi demi tujuannya tercapai. Menurut Pakpahan (2017), Teknologi Informasi untuk Indonesia sendiri ikut berkembang pesat didapatkan pengguna internet di Indonesia saat ini berjumlah 132,7 juta atau 52% dari jumlah penduduk Indonesia.

Penelitian terkait *hoax* pernah dilakukan oleh Petkovic et al, 2005, Vukovic et al, 2009, Chen et al.2014, Faisal Rahutomo et al.2019, namun penelitian tersebut

terkait email *hoax* dan sistem klasifikasi untuk berita *hoax* dengan menggunakan metode Naive Bayes, Fuzzy Logic, Neural Network, dan SVM . Kesimpulan pada penelitian analisis sentimen sebelumnya dilakukan oleh Pantouw (2017) mengenai Multinomial Naive Bayes pada pesan Twitter menunjukkan akurasi yang cukup baik yaitu 85.399%. Dan accuracy dari Faisal Rahutomo kurang baik yaitu rata- rata 82.6.

Penelitian terkait berita *hoax* dengan IndoBERT belum pernah dilakukan, oleh karena itu, penelitian ini bertujuan untuk mengklasifikasikan data berita *hoax* berbahasa Indonesia dengan metode IndoBERT dengan dataset yang didapat dari website pemerintah yaitu turnbackhoax.id yang selanjutnya dicrawling dan dijadikan dataset.

## 1.1 Rumusan Masalah

Dalam makalah ini akan dipaparkan beberapa poin :

1. Bagaimana masyarakat dapat memastikan tingkat kredibel berita-berita yang tersebar di internet dengan mudah ?

## 1.2 Tujuan

Penelitian yang dilakukan bertujuan :

1. Mengembangkan model yang dapat mengklasifikasikan berita *hoax*.

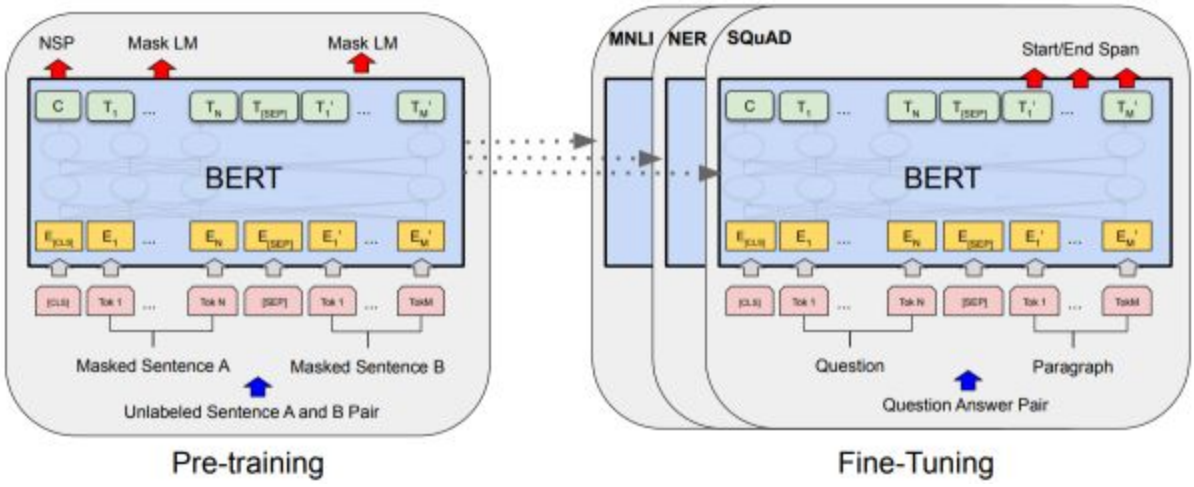
## 1.3 Manfaat

Manfaat yang didapat dari penelitian ini :

1. Hasil dari analisis ini dapat dipertimbangkan kembali untuk deploy model *deep learning* untuk permasalahan *hoax*
2. Pemerintah dapat mengembangkan kembali model *hoax* ini agar lebih terpercaya dan terjamin.
3. Permasalahan *hoax* pada masyarakat dapat menurun secara signifikan.

# 2. Metode

Metode yang dipakai dalam penelitian ini adalah menggunakan metode IndoBERT (*Indonesia Bidirectional Transformers for Language Understanding*). Metode BERT dipopulerkan oleh Google pada 11 Oktober 2018 tahun lalu dan revisi terakhir dilakukan pada 24 Mei 2019. BERT adalah model representasi bahasa baru yang menghasilkan model *pre-train* representasi *bidirectional* dari teks yang tidak berlabel dengan bersama-sama mengkondisikan dari kedua konteks di semua layer. Sehingga, model BERT yang telah dilatih, dapat disesuaikan dengan hanya menambahkan satu layer output saja.




Gambar 2.1 Langkah-Langkah yang Digunakan pada Metode BERT

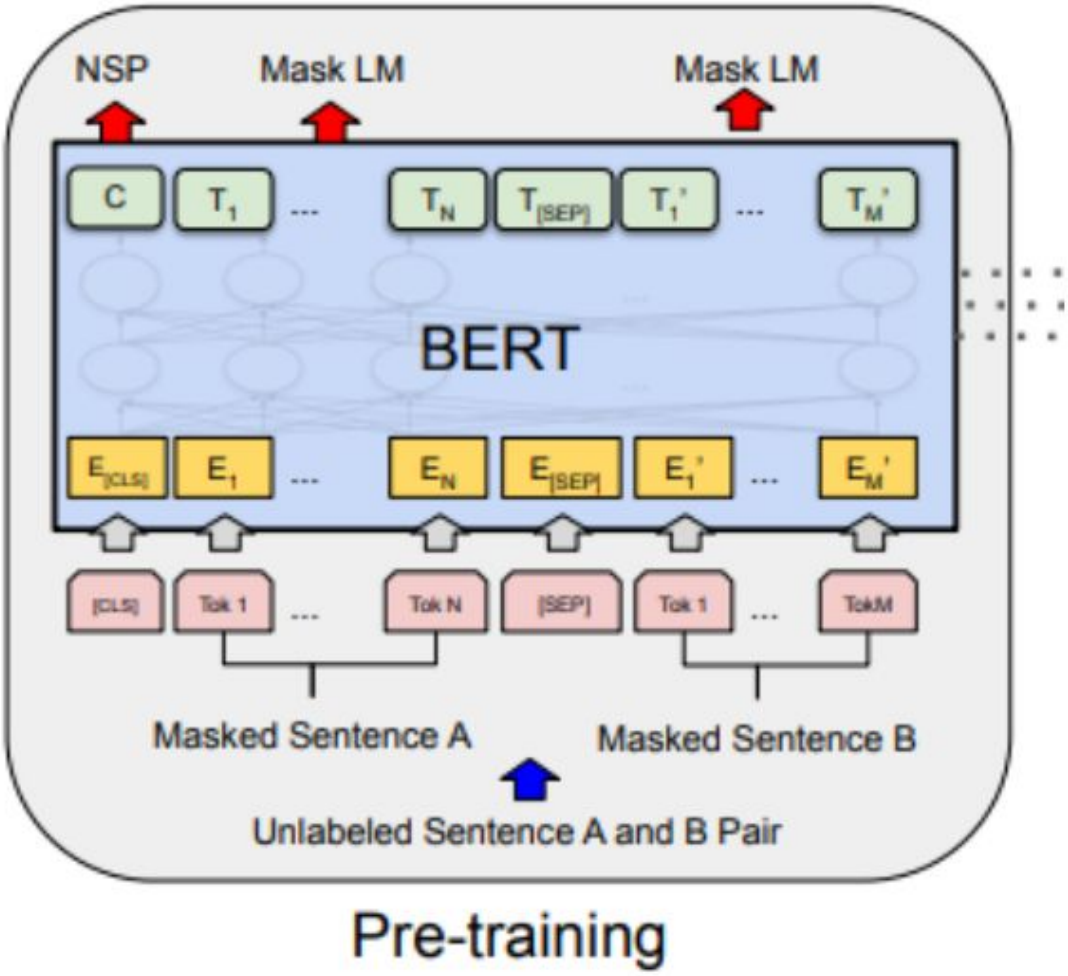
Seperti yang terlihat pada Gambar 2.1, terdapat 2 steps penting yang digunakan dalam metode ini, yaitu *pre-train* dan *fine-tuning*. Dalam step *pre-train*, dilakukan training model terhadap data yang tidak berlabel, sedangkan dalam step *fine-tuning* model BERT diinisialisasi dengan parameter pre-train, dan semua parameter tersebut disesuaikan dengan menggunakan data yang berlabel.

## 2.1 IndoNLU: IndoBERT

Indonesia pada tahun 2020 tepatnya tanggal 8 Oktober, Seperti yang telah diketahui Indonesia dikenal sebagai yang keempat dengan bahasa yang paling sering digunakan di internet, kemajuan penelitian tentang bahasa ini dalam pemrosesan bahasa alami (NLP) bergerak lambat karena kurangnya sumber daya yang tersedia. Namun perkembangan teknologi memaksakan penggunaan bahasa indonesia pada NLP dikembangkan, munculah IndoNLU mencakup dua belas tugas, mulai dari klasifikasi kalimat tunggal hingga urutan pasangan kalimat pelabelan dengan berbagai tingkat kerumitan. Kumpulan data untuk tugas terletak pada domain dan gaya yang berbeda untuk memastikan keragaman tugas. Model IndoBERT dilatih dari dataset yang bersih dan besar yaitu dataset Indonesia (Indo4B) dikumpulkan dari sumber yang tersedia untuk umum seperti teks media sosial, blog, berita, dan situs web.

## 2.2 Pre-train BERT

Untuk membuat representasi *bidirectional* yang baik dan akurat, maka perlu dilakukan *Masked Language Model* (MLM) yaitu suatu proses beberapa persen dari total token input diberi *'mask'* atau ditutup secara acak dan kemudian dicari prediksi terhadap token input yang sudah diberi *'mask'* tersebut.




Gambar 2.2 Langkah Pre-Train

Dalam Gambar 2.2 terdapat [CLS] yang merupakan token pertama dari setiap urutan yang dinamakan sebagai Special Classification Token ([CLS]) dan token terakhir merepresentasikan tugas klasifikasi. Simbol [SEP] pada gambar merupakan token khusus. Simbol C adalah vektor tersembunyi terakhir dari token [CLS], sedangkan simbol Ti adalah vektor tersembunyi terakhir untuk token input ke i. Pada step ini, vektor tersembunyi terakhir yang terkait dengan token *'mask'* dimasukkan kedalam softmax output melalui vocabulary, seperti pada *Language Modelling* yang sederhana.

## 2.3 Fine-tuning IndoBERT

Model *pre-train* yang dihasilkan memudahkan IndoBERT untuk memodelkan banyak *task* dengan cukup mengganti *input* dan *output* yang sesuai. IndoBERT menggunakan mekanisme *self-attention* untuk aplikasi yang melibatkan pasangan teks. Pada setiap *task, fine-tuning* cukup dilakukan dengan mencocokkan *input* dan *output* spesifik dan menyetel semua parameter secara *end-to-end*. Dalam melakukan *fine-tuning* diharuskan untuk memberikan dan mengkonfigurasi output layer dari pada arsitektur IndoBERT agar sesuai dengan output yang diinginkan.

‘

# 3.Desain dan Testing

## 3.1 Gambaran Sistem

Sistem yang dibangun memiliki 2 bagian**,** *Training* dan *Testing***.** Pada kedua bagian ini terdapat gambaran arsitektur yang digunakan untuk membangun model beserta penjelasannya.

### 3.1.1 Teknologi yang digunakan

#### 3.1.1.1 Tensorflow

Tensorflow adalah library dalam bahasa pemrograman python berlisensi *open source* yang menyediakan sarana untuk membangun model *deep learning.* Penggunaan *library* Tensorflow pada penelitian ini dikarenakan komunitas yang luas, kemudahan dalam membuat model, dan kemudahan dalam melakukan *debugging* pada saat membangun model .

#### 3.1.1.2 Pytypo

Typo adalah *library* dalam bahasa pemrograman python yang pada penelitian ini digunakan untuk membenarkan kata yang salah/*typo*.

#### 3.1.1.3 Transformers

Transformers adalah *library* dalam bahasa pemrograman python berlisensi *open source.* yang menyediakan model-model *pretrained* untuk melakukan suatu *tasks* pada teks. Penggunaan *library* transformers pada penelitian ini bertujuan untuk menggunakan model *pretrained* IndoBERT dan *preprocessing data* .

#### 3.1.1.4 Tensorflow addons

Tensorflow *addons* adalah library dalam bahasa pemrograman python berlisensi open source yang menyediakan tambahan fungsi yang tidak terdapat pada

*library core Tensorflow*. *Library Tensorflow addons* yang digunakan pada penelitian ini, untuk mengatasi data yang tidak seimbang.

3.1.1.5 Pandas

Pandas adalah *library* dalam bahasa pemrograman python berlisensi *open source* yang menyediakan struktur data dan analisis data. *library* Pandas yang digunakan dalam penelitian ini untuk membuat *data frame* dan memanipulasi *data frame*.

3.1.1.6 Numpy

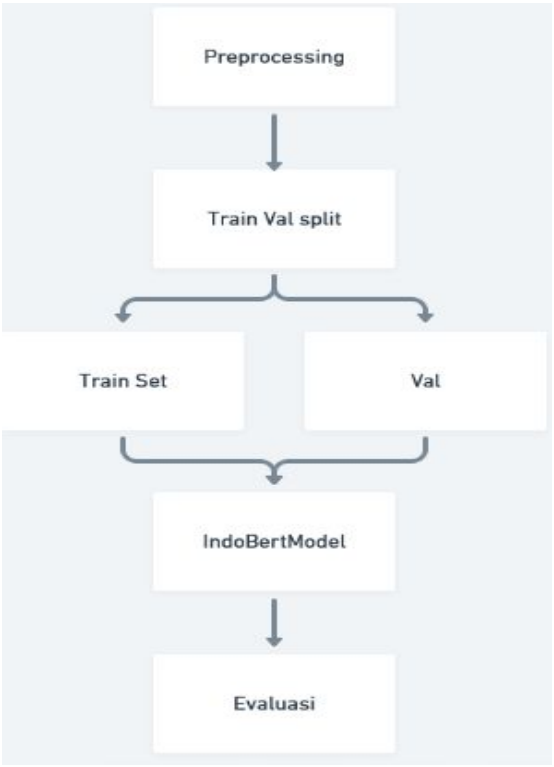
Numpy adalah *library* dalam bahasa pemrograman python berlisensi *open source* yang pada penelitian ini numpy digunakan untuk mencari *index* pada *array* yang memiliki nilai maksimal.

3.1.1.7 Emoji

Emoji adalah *library* dalam bahasa pemrograman python berlisensi open source yang menyediakan *module* untuk menghilangkan *emoji* pada suatu teks. Penggunaan *library* Emoji bertujuan untuk untuk mempercepat *preprocessing emoji* pada teks.

### 3.1.2 **Training**

Proses *training* dalam *deep learning* melewati beberapa proses penting yang harus dilakukan ,data yang bersih, memecah dataset, training, dan evaluasi. Berikut *Flowchart* metode yang digunakan .




Gambar 3.1.2 ( Training Dengan IndoBERT )

### 3.1.3 Testing


Gambar 3.1.3 ( Testing )

## 3.2 *Exploratory Data Analysis* (EDA)

*Exploratory Data Analysis* (EDA) adalah proses menggali intuisi (memahami) terhadap data yang digunakan, apakah data ini dan untuk apa data ini[6]. Dataset yang digunakan telah dikumpulkan secara mandiri dari data *hoax* dan *non hoax* pada website turnbackhoax.id yang telah dijadikan dataset.

### 3.2.1 Anomali Data

Dari hasil analisa data ditemukan kata yang sering muncul atau disebut juga *most common words.* Berikut disajikan *Word Cloud* untuk menggambarkan banyaknya kemunculan suatu kata.


Gambar 3.2.1a (kata yang sering muncul dari judul yang bersifat hoax)


Gambar 3.2.1b (kata yang sering muncul dari judul yang bersifat non-hoax)


Gambar 3.2.1c (kata yang sering muncul dari narasi yang bersifat hoax)


Gambar 3.2.1d (kata yang sering muncul dari narasi yang bersifat non-hoax)

## 3.3 Training

Tahap pertama dilakukan proses *fine-tuning* dengan model IndoBERT untuk studi kasus ini. *Framework* Tensorflow 2.0 digunakan untuk memodifikasi model, dan *library* Transformers untuk menggunakan *pretrained model*. Kode implementasi yang digunakan untuk training disini.

### 3.3.1 Preprocessing Dataset

Hasil pengambilan judul dan narasi berita kotor dan tidak baik untuk inputan model dengan demikian akan dilakukan tahap *Preprocessing* terlebih dahulu, dengan langkah langkah sebagai berikut :

1. Menjadikan kata kata ke huruf kecil

   Mempermudah dalam pengubahan singkatan menjadi bentuk kata aslinya

2. Menghapus simbol-simbol yang tidak perlu

   Mempermudah proses *training* dan *attentions*

3. Mengubah *link menjadi* kata lain

   Pengambilan data melalui turnbackhoax.id terkadang terdapat *link/hyperlink* yang dapat merugikan fitur dalam *train* model, jadi diganti dengan kata "alamat web".

4. Menghapus spasi tambahan

   Terdapat spasi tambahan dapat menjadi permasalahan dalam pemrosesan.

5. Mengambil 36 karakter utama untuk judul dan 96 karakter utama untuk narasi

   Membatasi sebuah inputan, agar model tetap seimbang.

6. Menghilangkan *emoji*

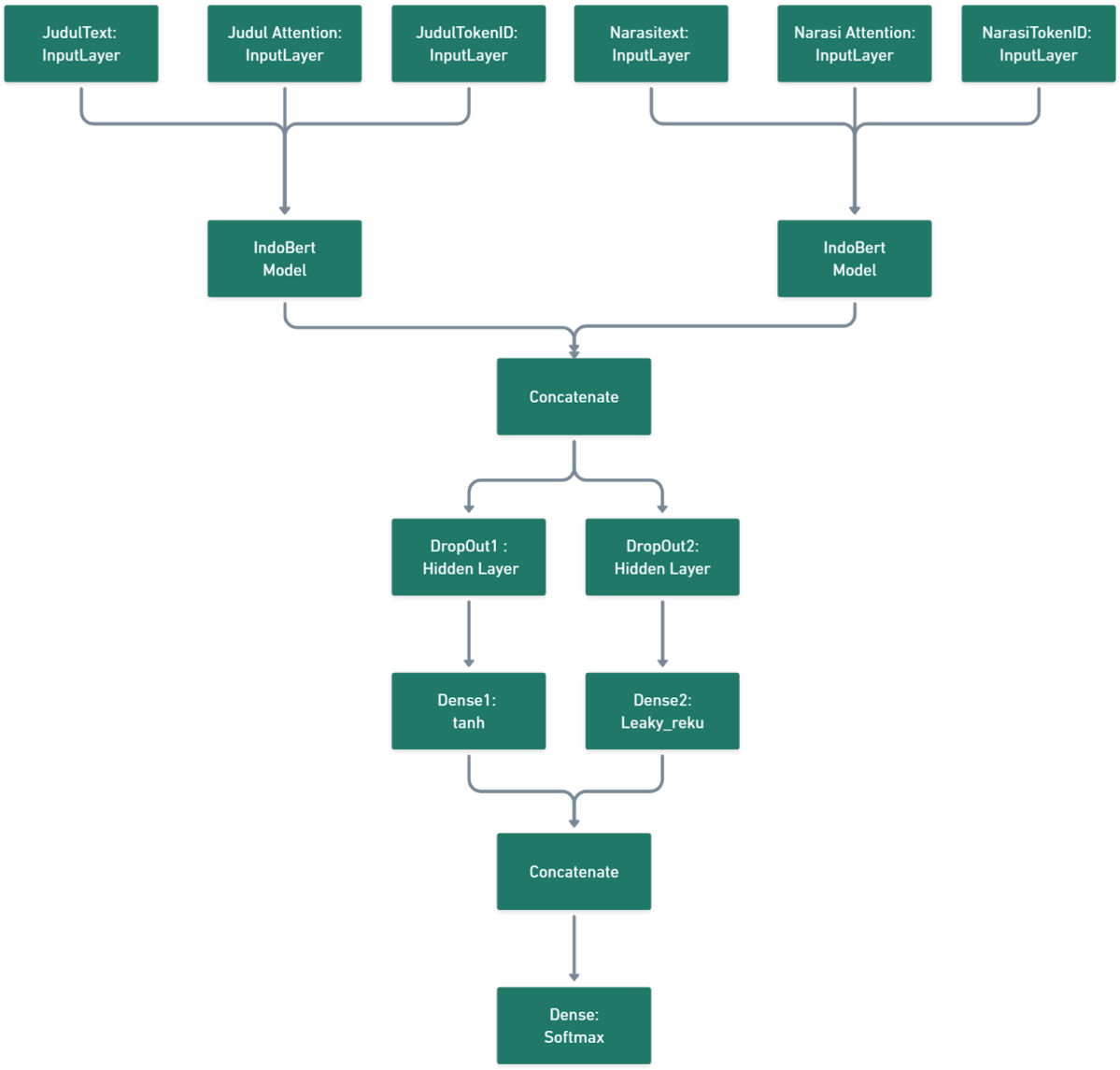
*Emoji* biasa digunakan untuk mengekspresikan sesuatu, untuk kasus ini akan dihilangkan.

7. Encode teks berita menjadi inputan ke model

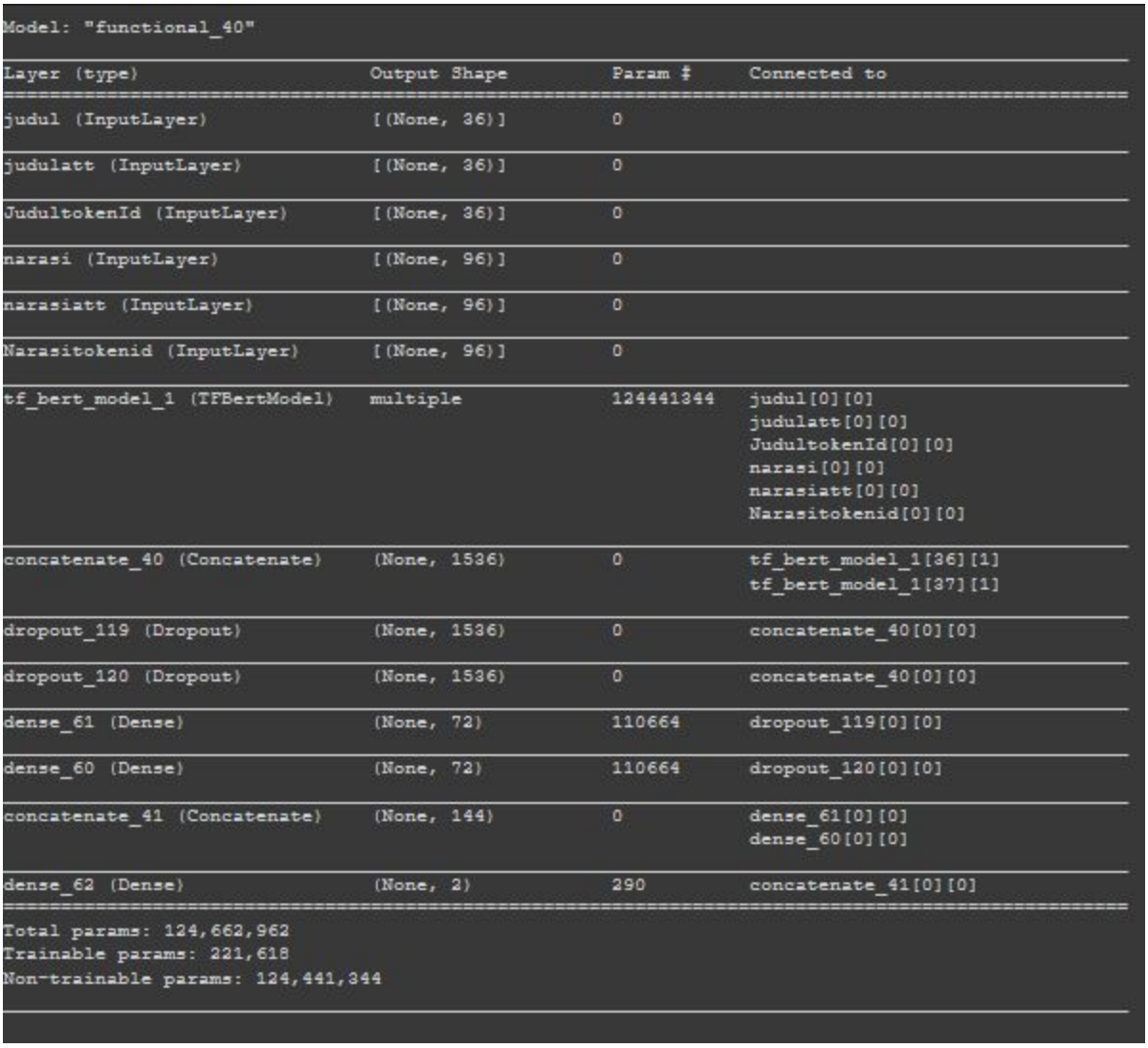
Preprocessing akhir adalah mengubah seluruh kata yang ada dengan inputan yang dibutuhkan oleh model IndoBERT .

### 3.3.3 Arsitektur Jaringan Neural Network

IndoBERT sendiri adalah arsitektur *pre-training* sehingga tidak dapat digunakan langsung terhadap kasus tertentu, dalam kasus ini dilakukan adalah "*Sentiment analysis"*. Oleh karena itu output layer dari IndoBERT harus dibentuk lagi dengan menambahkan layer baru untuk menyelesaikan kasus ini. Berikut arsitektur yang telah dibuat.




Gambar 3.3.3.1 (Arsitektur Model)



```
Model: "functional_40"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
judul (InputLayer)              [(None, 36)]         0
__________________________________________________________________________________________________
judulatt (InputLayer)           [(None, 36)]         0
__________________________________________________________________________________________________
JudultokenId (InputLayer)       [(None, 36)]         0
__________________________________________________________________________________________________
narasi (InputLayer)             [(None, 96)]         0
__________________________________________________________________________________________________
narasiatt (InputLayer)          [(None, 96)]         0
__________________________________________________________________________________________________
Narasitokenid (InputLayer)      [(None, 96)]         0
__________________________________________________________________________________________________
tf_bert_model_1 (TFBertModel)   multiple             124441344   judul[0][0]
                                                                 judulatt[0][0]
                                                                 JudultokenId[0][0]
                                                                 narasi[0][0]
                                                                 narasiatt[0][0]
                                                                 Narasitokenid[0][0]
__________________________________________________________________________________________________
concatenate_40 (Concatenate)    (None, 1536)         0           tf_bert_model_1[36][1]
                                                                 tf_bert_model_1[37][1]
__________________________________________________________________________________________________
dropout_119 (Dropout)           (None, 1536)         0           concatenate_40[0][0]
__________________________________________________________________________________________________
dropout_120 (Dropout)           (None, 1536)         0           concatenate_40[0][0]
__________________________________________________________________________________________________
dense_61 (Dense)                (None, 72)           110664      dropout_119[0][0]
__________________________________________________________________________________________________
dense_60 (Dense)                (None, 72)           110664      dropout_120[0][0]
__________________________________________________________________________________________________
concatenate_41 (Concatenate)    (None, 144)          0           dense_61[0][0]
                                                                 dense_60[0][0]
__________________________________________________________________________________________________
dense_62 (Dense)                (None, 2)            290         concatenate_41[0][0]
==================================================================================================
Total params: 124,662,962
Trainable params: 221,618
Non-trainable params: 124,441,344
__________________________________________________________________________________________________
```

Gambar 3.3.3.2 (Summary Model)

Setelah Inputan di proses pada Model IndoBERT ditambahkan 1 *layer* yaitu *layer concatenate* untuk menggabungkan *output* dari inputan judul dan inputan narasi yang selanjutnya *output* dari *concatenate* tersebut akan masuk kepada 2 *layer dropout* agar tidak terjadi *overfitting* pada model, lalu setelah *output* pada *dropout* akan masuk ke *layer dense* dengan masing-masing 72 *neurons* dengan *activation leaky relu* untuk menghindari *Dead Relu* dan pada salah satu *dense layer* dan *tanh* untuk menghindari bias pada *gradients*. *Output dense* masing-masing akan masuk ke *layer concatenate* dan terakhir akan masuk ke *layer Dense* dengan *softmax activation* dengan 2 *units*

untuk *output*. *loss function* yang digunakan adalah *"SigmoidFocalCrossEntropy"*. Digunakan karena data yang dimiliki imbalance .

Dalam proses training digunakan *Hyperparameter* sebagai berikut:

- Model IndoBERT : indobenchmark/indobert-base-p2
- Max Sequence Length : 36 Judul dan 96 Narasi
- Loss Function : **SigmoidFocalCrossEntropy**
- Optimizer : Adam
- Epoch : 15
- Validation Split : 0.05
- Learning Rate : Using learning rate scheduler
- Test Size : 470

Hasil dari train dan validation dataset

| Data Set | Jumlah Data | Akurasi |
|---|---|---|
| Train Set | 8733 | 0.8786 |
| Validation Set | 460 | 0.8978 |

Tabel 3.3.3.1 (Hasil Training)

## 3.4 Testing

### 3.4.1 Preprocessing Dataset

Hasil pengambilan judul dan narasi berita kotor dan tidak baik untuk inputan model dengan demikian akan dilakukan tahap *Preprocessing* terlebih dahulu, dengan langkah langkah sebagai berikut :

1. Menjadikan kata kata ke huruf kecil

   Mempermudah dalam pengubahan singkatan menjadi bentuk kata aslinya

2. Menghapus simbol simbol yang tidak perlu

   Mempermudah proses training dan attentions

3. Mengubah *link menjadi* kata lain

   Pengambilan data melalui turnbackhoax.id terkadang terdapat *link/hyperlink* yang dapat merugikan fitur dalam train model, sehingga diganti dengan kata “alamat web”.

4. Menghapus spasi tambahan

   Terdapat spasi tambahan, yang dapat menjadi permasalahan pada training.

5. Mengambil 36 karakter utama untuk judul dan 96 karakter utama untuk narasi

   Membatasi sebuah inputan, agar model tetap seimbang.

6. Menghilangkan emoji

   Emoji biasa digunakan untuk mengekspresikan sesuatu, untuk kasus ini emoji dihilangkan.

7. *Encode* teks berita menjadi inputan ke model

*Preprocessing* akhir adalah mengubah seluruh kata yang ada dengan inputan yang dibutuhkan oleh model IndoBERT .

Hasil dari pada Test set

| Data Set | Jumlah Data | F1-Score | Precision | Recall |
|---|---|---|---|---|
| Test Set | 470 | 0.92 | 0.87 | 0.72 |

Tabel 3.3.3.2 (Hasil Test set)

# 4. Analisis

## 4.1 Analisa Data

*Dataset* yang berukuran 470 ini didapatkan dari turnbackhoax.id. Model Memprediksi Seluruh dataset dan didapat hasil daripada prediksi test set ini, terdapat beberapa misklasifikasi, berikut penjelasan analisa luaran model.

| No | Judul | Narasi | Label | Prediction | Justifikasi Luaran Model |
|---|---|---|---|---|---|
| 1. | UANG NKRI EDISI 2016 BERHASIL DIPALSUKAN | Heboh, belum genap setahun, uang NKRI Edisi 2016 ini berhasil Dipalsukan! Begini ciri-cirinya | 1 (*HOAX*) | 1 (*HOAX*) | Model memprediksikan seperti ini karena,terdapat kesesuaian antara judul dan narasi |
| 2. | Kominfo Bantah Adanya Kebocoran Anggaran Asian Games Sebesar Rp. 846 Juta di Asian Games | Pernyataan saudara Jajang Nurjaman terhadap potensi kerugian negara yang membandingkan harga penawaran PT Indo-Ad dengan PT Bee Work Pariwara tidak relevan, ujar Plt Kepala Biro Humas Kominfo, Noor Iza, Selasa (21/8) | 0 (*Non HOAX*) | 0 (*Non HOAX*) | Model memprediksikan seperti ini karena didapatkan sebuah pernyataan dari seorang narasumber yang kredibel dan adanya kesesuaian antara judul dan narasi . |
| 3. | PT KCI | Permintaan maaf | 0 | 1 (*HOAX*) | Model |

| | | | | | |
|---|---|---|---|---|---|
| | Mengklarifikasi Kabar Tiket Kertas Gratis di Stasiun Bogor Pada Senin 23 Juli 2018 | khususnya kami sampaikan kepada para pelanggan setia kami | | | memprediksikan seperti ini karena berita yang belum cukup jelas dan hubungan antara judul dan narasi pun tidak didapatkan |
| 4 | Masih Banyak yang Baik dari Internet | Masih Banyak yang Baik dari Internet | 0(*Non HOAX*) | 1 (*HOAX*) | Model memprediksikan seperti ini karena berita yang belum cukup jelas |
| 5 | PLN Bantah Informasi Tenggat Pembayaran Listrik Dimajukan Tanggal 5 Setiap Bulannya | Sekadar info, mulai bln Maret 2019 pembayaran tagihan PLN dimajukan. Biasanya paling lambat tgl 20 tiap bulan, dimajukan ke tgl 5. Pembayaran sesudah tgl 5 sdh kena denda. Mohon bantu share kpd keluarga, tetangga, teman2. Indahnya berbagi | 0(*Non HOAX*) | 1 (*HOAX*) | Model memprediksikan seperti ini karena tidak terdapat pemaparan yang jelas daripada narasi yang disampaikan dan adanya ketidaksesuaian antara judul dan narasi. |
| 6 | Klarifikasi KPK Terkait Foto | Secara tidak sengaja sekitar | 0(*Non HOAX*) | 1 (*HOAX*) | Model memprediksikan |

| | | | | | |
|---|---|---|---|---|---|
| | Setya Novanto Tanpa Baju Tahanan | pukul 06.00 WIB di rest area kilometer 97 Tol Purbaleunyi arah Jakarta, rombongan Investigasi Gabungnya Wartawan Indonesia (GWI) melihat orang dengan pengawalan mirip mantan Ketua DPR-RI, Setya Novanto (Setnov) terpidana 15 tahun penjara ! | | | seperti ini karena tidak ada pernyataan yang kredibel lalu ada kata “mirip” dan ketidaksesuaian judul dan narasi. |
| 7 | Idola 92.6 FM Semarang: Merefleksi Hari Media Sosial di Tengah Pandemi Covid-19 | Ngobrol-ngobrol Idola 92.6 FM Semarang dengan Aribowo Sasmito, Ketua Komite Pemeriksa Fakta MAFINDO, berkaitan dengan Hari Media Sosial (Rabu, 10 Jun 2020). Simak di: | 0(Non *HOAX*) | 1 (*HOAX*) | Model memprediksikan seperti ini karena kurangnya informasi lebih detail terkait narasai |

Tabel 4.1.1 (Hasil Analisa test set)

# 5. Kesimpulan

Penggunaan Pretrained Model IndoBERT dapat menunjukan hasil yang signifikan pada suatu permasalahan dengan dataset yang tidak banyak, hal ini terbukti model dapat meraih *accuracy test set* sebesar 92%. Tentu ada banyak sisi yang harus dipertimbangkan kembali agar model yang dihasilkan menggeneralisasi lebih baik lagi, seperti menambah data, mengetahui asal data tersebut, preprocessing data dengan benar, penggunaan fungsi aktivasi, fungsi *loss*, fungsi *optimizer*, dan perangkaian arsitektur yang tepat berpengaruh terhadap *output* yang dihasilkan model.

Hasil model yang didapatkan memprediksi seluruh berita dengan baik, model ini dapat dipertimbangkan untuk penggunaan secara live guna mengurangi berita *hoax* yang beredar di masyarakat dan keluaran klasifikasi dari model dapat dijadikan sebagai acuan bagi masyarakat umum untuk menilai kredibilitas suatu berita yang akan ia baca. Permasalahan infrastruktur, pengetahuan teknologi, ekonomi, dan sosial perlu juga dibenahi agar permasalahan *hoax* di Indonesia dapat teratasi.

# DAFTAR PUSTAKA

[1]kumparanTECH. “Riset: 64% Penduduk Indonesia Sudah Pakai Internet.” *Kumparan*, Kumparan, 21 Feb. 2020, kumparan.com/kumparantech/riset-64-penduduk-indonesia-sudah-pakai-internet-1ssUCDbKILp[Diakses 14 11 2020].

[2]Kominfo, Pdsi. "Ada 800.000 Situs Penyebar Hoax Di Indonesia." Website Resmi Kementerian Komunikasi Dan Informatika RI. https://kominfo.go.id/content/detail/12008/ada-800000-situs-penyebar-hoax-di-indonesia/0/sorotan_media [Diakses 14 11 2020].

Faisal Rahutomo, Inggrid Yanuar Risca Pratiwi, Diana Mayangsari Ramadhani “EKSPERIMEN NAÏVE BAYES PADA DETEKSI BERITA HOAX BERBAHASA INDONESIA “ (2019).

Wilie, B dkk. (2020). IndoNLU: *Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. V1.