**DOKUMENTASI PROJECT**

**PEMROSESAN BAHASA ALAMI**

**DETEKSI BERITA VALID dan NON VALID (HOAX)**



**NAMA ANGGOTA :**

1. Ahmad Junjung Sudrajad (201910370311405)
2. Aris Muhandisin (201910370311432)
3. Lidya Fankky Oktavia Putri (201910370311448)

**PROGRAM STUDI INFORMATIKA**

**FAKULTAS TEKNIK**

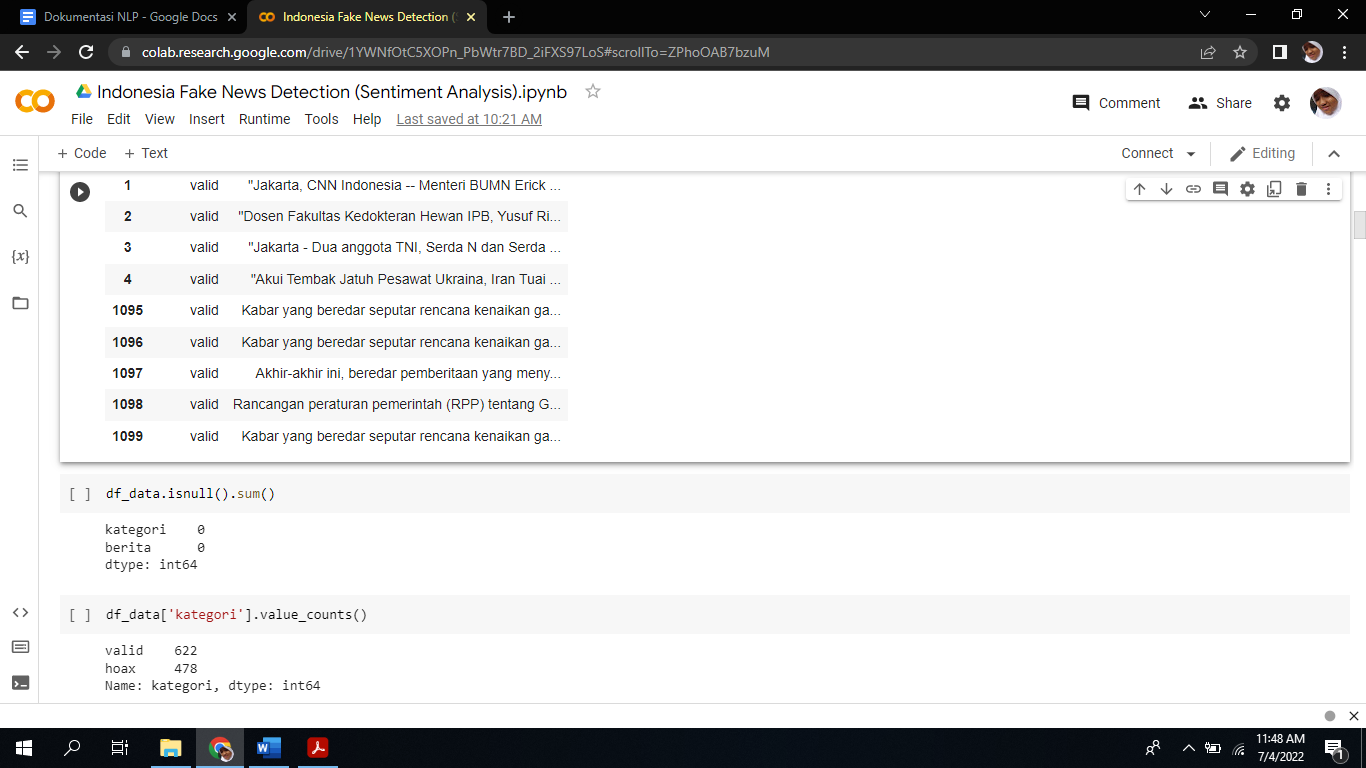
**UNIVERSITAS MUHAMMADIYAH MALANG**

**2022**

## **Tentang Project**

Project yang dikerjakan pada penelitian ini merupakan deteksi berita valid dan non valid. Berita non valid selalu menjadi masalah di berbagai belahan dunia. Semenjak bahasa Inggris adalah bahasa yang paling dominan di dunia, penganalisis berita non valid sebagian besar dibuat untuk memenuhi atau menganalisa berita yang dituliskan dalam bahasa Inggris. Dengan begitu, pada penelitian kali ini menggunakan data berita yang berbahasa Indonesia. Data tersebut akan dilakukan pemrosesan analisis menggunakan teknologi informasi yaitu analisis sentimen.

## **Dataset**

Dataset yang digunakan merupakan hasil dari jurnal referensi dan hasil crawling sendiri. Kemudian dataset tersebut digabung menjadi satu, sehingga total keseluruhan dari dataset yaitu 1099. Dataset tersebut dibagi sebanyak 2 kategori yaitu berita valid dan berita hoax (non valid). 

## **Data Pre-processing**

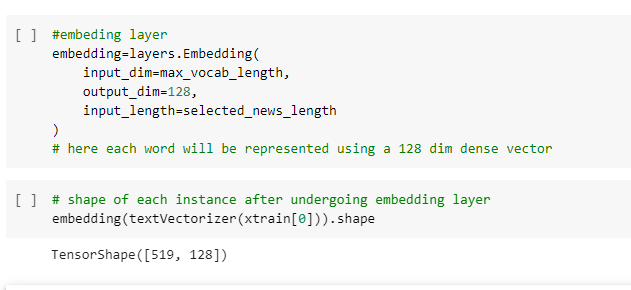
* Pembagian dataset menjadi 2 direktori yakni train dan test 8:2
* Tokenization dan Vectorization

Yang bertujuan memberi nilai int pada dataset yang masih berupa text sehingga bisa diolah pada proses pelatihan model.



* Embedding Layer

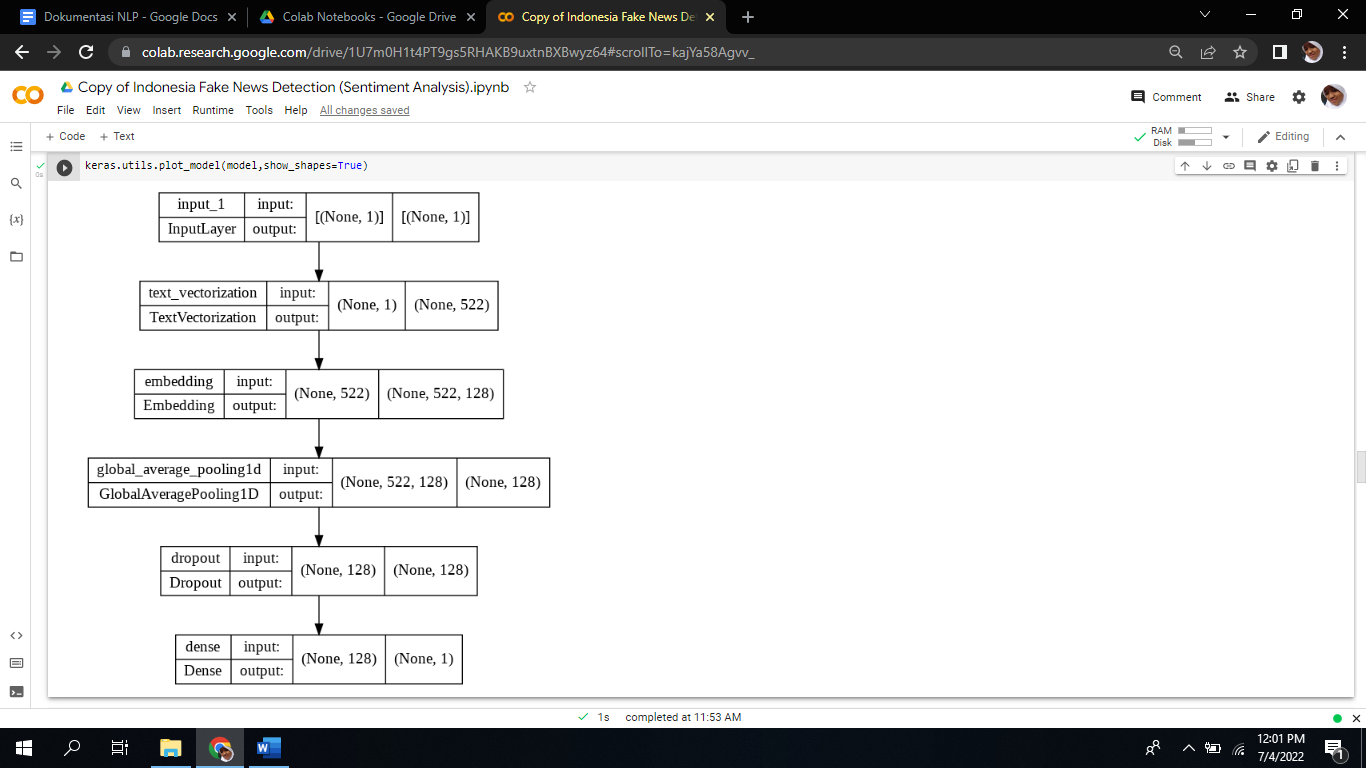
Merupakan suatu step yang bertujuan untuk memberikan bobot pada layer yang akan dibangun pada model dan data dengan parameter output\_dim sebesar 128



## **Model**

## Dalam rancangan Arsitektur model yang akan penulis rancang menggunakan beberapa model. Arsitektur Model yang akan penulis gunakan adalah :

1. Model FNN

Arsitektur model FNN dapat dilihat seperti gambar dibawah ini :

Penggunaan model FNN dalam studi kasus analisis sentimen berita valid dan hoax (non valid) menghasilkan nilai akurasi sebesar 81%.

1. Naiv Bayes (MultinomialNB) + TFIDFVectorizer

Dalam analisis sentimen berita valid dan hoax (non valid) yang dilatih menggunakan model Naiv Bayes (MultinomialNB) + TFIDFVectorizer menghasilkan nilai akurasi sebesar 70%.

## **Visualisasi**

1. Visualisasi dari WordCloud

Dalam hal ini penulis membuat dua WordCloud yaitu pada label valid dan label hoax (non valid), sehingga bisa mengetahui kata apa saja yang sering muncul dalam setiap label, Untuk Source Code sebagai berikut:

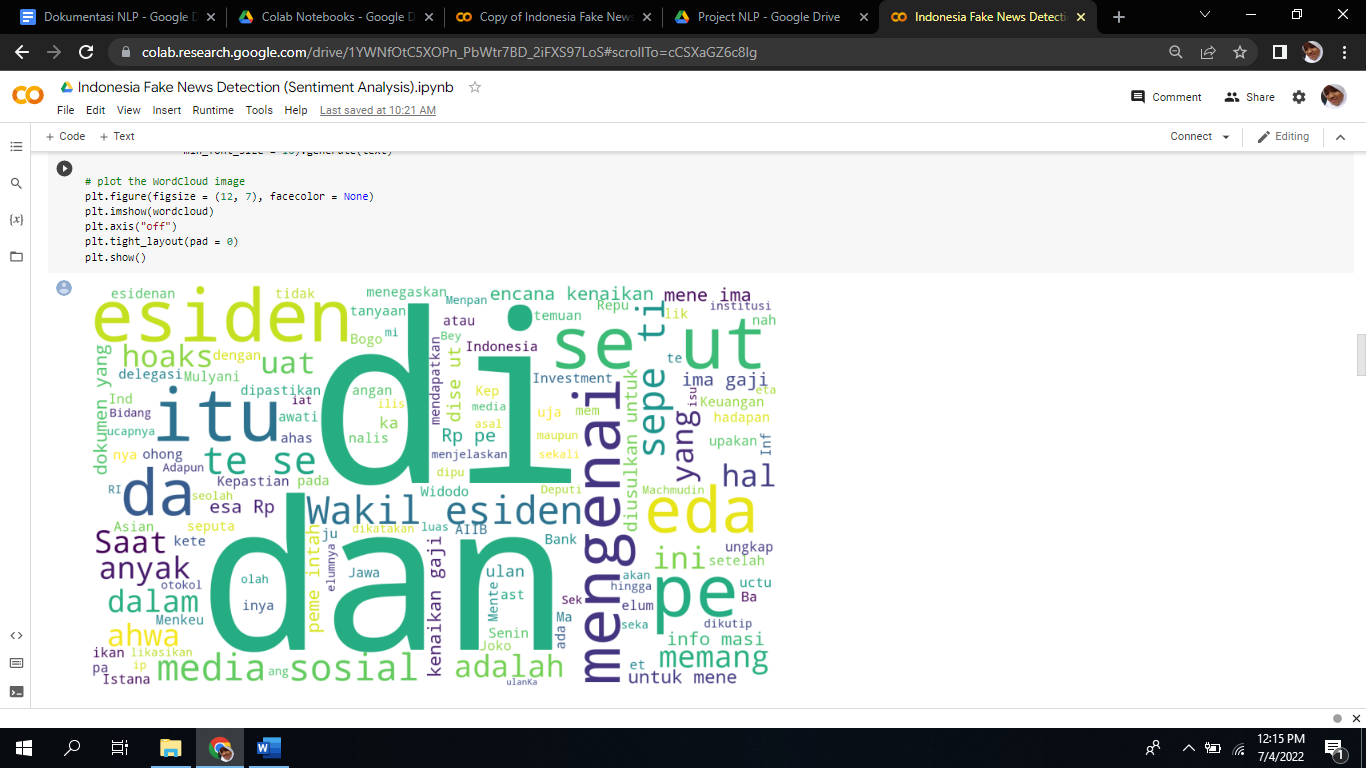
* Valid rating WordCloud

| text\_list=df\_data.loc[df\_data['kategori']==0,'berita'].tolist()  text=""  for sent in text\_list:  sent=re.sub("[<br>]"," ",sent)  text=re.sub("[^A-Za-z0-9]+"," ",sent)  text+=sent  stopwords=set(STOPWORDS)  wordcloud = WordCloud(width=1200,height=700,  background\_color ='white',  stopwords = stopwords,  min\_font\_size = 10).generate(text)    # plot the WordCloud image  plt.figure(figsize = (12, 7), facecolor = None)  plt.imshow(wordcloud)  plt.axis("off")  plt.tight\_layout(pad = 0)  plt.show() |
| --- |

* Hoax rating WordCloud

| text\_list=df\_data.loc[df\_data['kategori']==1,'berita'].tolist()  text=""  for sent in text\_list:  sent=re.sub("[<br>]"," ",sent)  text=re.sub("[^A-Za-z0-9]+"," ",sent)  text+=sent  stopwords=set(STOPWORDS)  wordcloud = WordCloud(width=1200,height=700,  background\_color ='white',  stopwords = stopwords,  min\_font\_size = 10).generate(text)    # plot the WordCloud image  plt.figure(figsize = (12, 7), facecolor = None)  plt.imshow(wordcloud)  plt.axis("off")  plt.tight\_layout(pad = 0)  plt.show() |
| --- |

Hasil dari visualisasi WordCloud dapat dilihat seperti gambar dibawah ini :

* Valid rating WordCloud

* Hoax rating WordCloud