

Twitter Preprocessing

L Tarangga Arief G



Kenapa Perlu Preprocessing?

Menghapus data yang tidak relevan

kebanyakan data teks yang diproleh penuh dengan kata atau symbol yang tidak memiliki makna

Teknik

Yang akan dibahas



- 01**
case folding
- 02**
Pembersihan URL, Mention, Hastag, reserve words, EMOJI, Smiley, Number
- 03**
Menghapus tanda baca
- 04**
Spelling Correction
- 05**
Pembersihan Stopwords

Makan

makan

text.lower()

membuat text menjadi huruf kecil semua



**vaksin yang diadakan
pemerintah sedikit
terhambah @presidenri**

! pip install tweet-preprocessor

Pembersihan URL, Mention, Hastag, reserve words,
EMOJI, Smiley, Number



Pilihan penghapusan

Option Name	Option Short Code
URL	p.OPT.URL
Mention	p.OPT.MENTION
Hashtag	p.OPT.HASHTAG
Reserved Words	p.OPT.RESERVED
Emoji	p.OPT.EMOJI
Smiley	p.OPT.SMILEY
Number	p.OPT.NUMBER

saya,

saya

```
re.sub(r'^\w\s]', "", text)
```

menghapus tanda baca

manusia sering error!

kadang kita di twitter sering menggunakan typo atau teks yang sengaja disingkat, nah ini membuat mesin kebingungan mana kata yang sama dan kata yang tidak sama



https://github.com/meisaputri21/Indonesian-Twitter-Emotion-Dataset/blob/master/kamus_singkatan.csv

Spelling Correction

**Dia sedang
makan di rumah**

**Dia makan di
rumah**

**! pip install sastrawi
atau
! pip install spacy**

Menghapus Stopword

stopword adalah kata-kata yang jika dihapus tidak akan mengubah makna di dalam teks



urutannya
bagaimana?

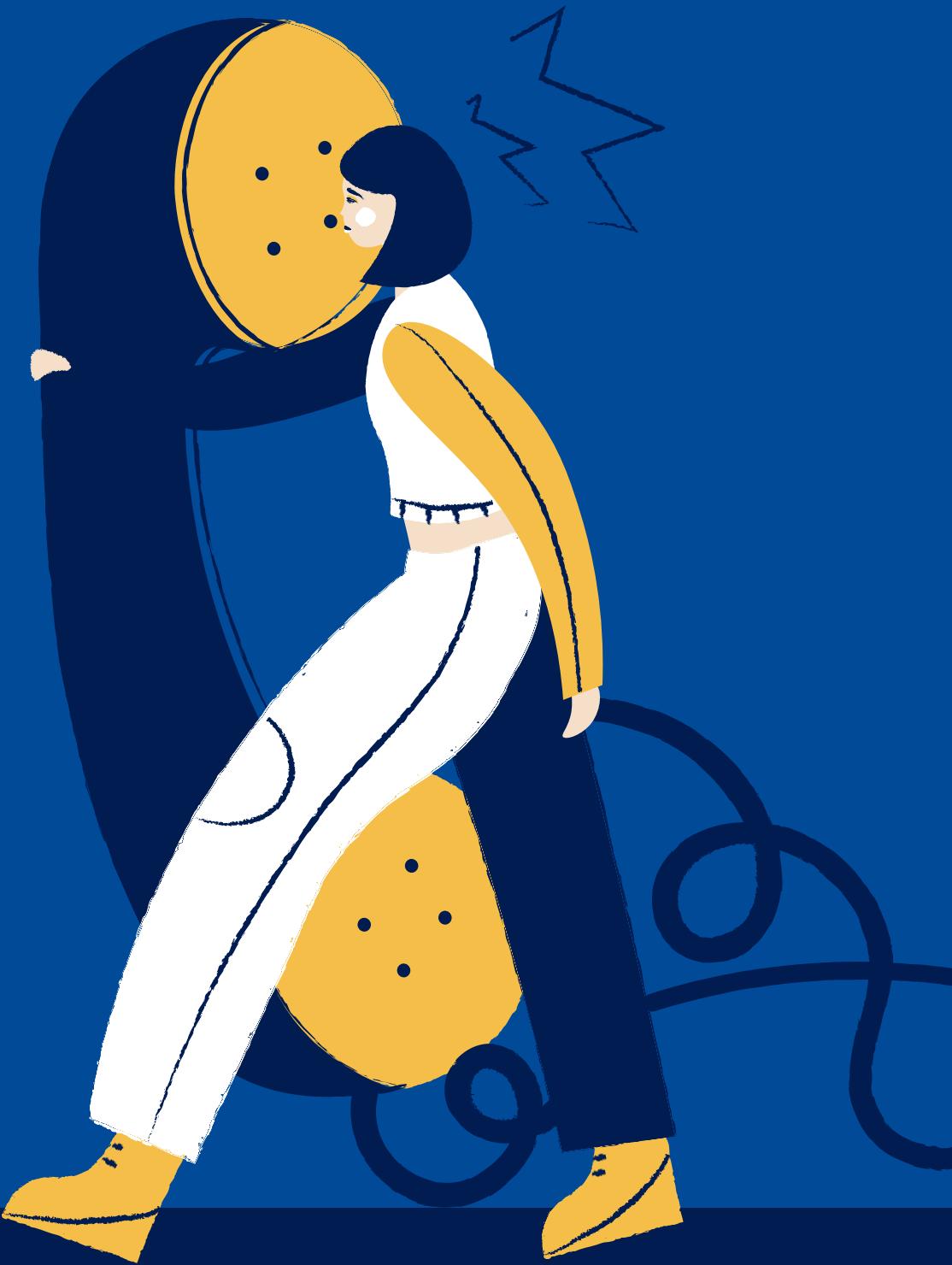
Terimakasih

Alamat Surel

tarangga.arief@gmail.com

Nomor Telepon

085339383968



Tambahan

<https://huggingface.co/>

daftar deep learning yang bisa langsung digunakan