Text Mining - Final Project

MACHINE LEARNING UNTUK KLASIFIKASI PESAN SPAM BAHASA INDONESIA

Menggunakan Naïve Bayes (TF-IDF)

Presented by:
Zaima Syarifa Asshafa (2110631250066)

Methodology

Workflow Diagram

Business Understanding

Identifikasi masalah, penentuan tujuan, dan matriks evaluasi.

Data Understanding

Pemahaman karakteristik dan kebutuhan data.

Exploratory Data Analysis

Penggalian mendalam terkait informasi pada data.

Data Preprocessing

Pemrosesan dan penanganan masalah pada data.

Modeling & Evaluation

Pembangunan model dengan beberapa algoritma dan evaluasi model.

Testing Model

Pengetesan klasifikasi teks dengan final model.

Business Understanding

What the problem?

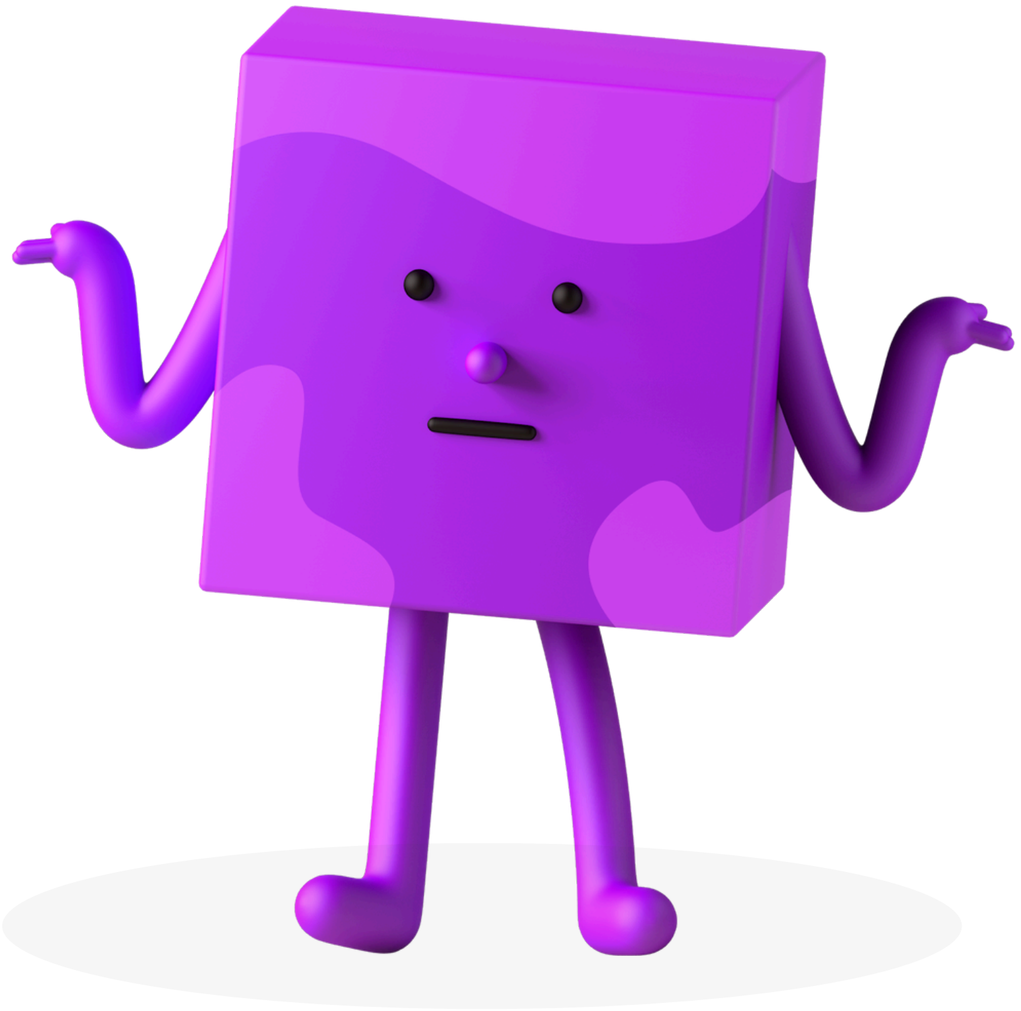
Bagaimana membedakan pesan normal/spam/promo berdasarkan teks yang terkandung dalam pesan?

Goals

Mengembangkan model klasifikasi pesan untuk mendeteksi spam berdasarkan teks.

Metrics Evaluation

Precision, recall, dan F1-score, dengan fokus pada **minimalisasi false positive** (pesan non-spam dianggap spam).



Data Understanding

Data Overview

	Teks	label
574	Di kfc yg deket enhaii ada dy	0
575	Maaf jika ada janji yang belum terpenuhi, jik...	0
576	*ngsih bunga ato coklat min	0
577	.sambl nunggu itu.. Gimana kalo ngerjain form ...	0
578	[Akademik] Untuk perhatian tuk jadwal kontrak ...	0
...	...	...
1137	Yooo sama2, oke nanti aku umumin di grup kelas	0
1138	sebelumnya ga ad nulis kerudung. Kirain warn...	0
1139	Mba mau kirim 300 ya	0
1140	nama1 beaok bwrangkat pagi...mau cas atay tra...	0
1141	No bri atas nama kamu mana	0

Pesan Normal

	Teks	label
239	Jika anda bermasalah dgn CC/KT@, stres dgn bun...	1
240	Lelah byr min payment? Kami Solusinya, bantu s...	1
241	Dana Tunai (KTA) bunga 0,99% hingga 300 jt. Sy...	1
242	“ROXI CELL” Hanya dengan Rp.100rb Anda bisa ja...	1
243	3 RAMADHAN Selamat Anda Pemenang Rp.100jt. PIN...	1
...	...	...
569	Yth Bpk/Ibu. BNI menyatakan Rekening anda terp...	1
570	Yth Isti Sofiyah. Diminta Segera Hubungi Bpk D...	1
571	YTH kpd bpk/ibu sy Eka Novitasari kariyawan 3c...	1
572	YTH,MIitra Silahkan cek poin anda dan tukarkan...	1
573	Yuk Ikuti akun dakwah, caranya: ketik IKUT [sp...	1

Pesan Spam

	Teks	label
0	[PROMO] Beli paket Flash mulai 1GB di MY TELKO...	2
1	2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat A...	2
2	2016-07-08 11:47:11.Plg Yth, sisa kuota Flash ...	2
3	2016-08-07 11:29:47.Plg Yth, sisa kuota Flash ...	2
4	4.5GB/30 hari hanya Rp 55 Ribu Spesial buat an...	2
...	...	...
234	Yuk INTERNET-an NGEBUT utk akses FB, Twitter, ...	2
235	Yuk temen belanja di google play, mudah banget...	2
236	Yuk tetap gunakan Flash Volume Ultima utk upda...	2
237	Mau nonton bioskop gratis bersama keluarga? Ci...	2
238	Yuks internetan seru-seruan dg Flash Volume Ul...	2

Pesan Promosi

Data Understanding

Import Necesary Library

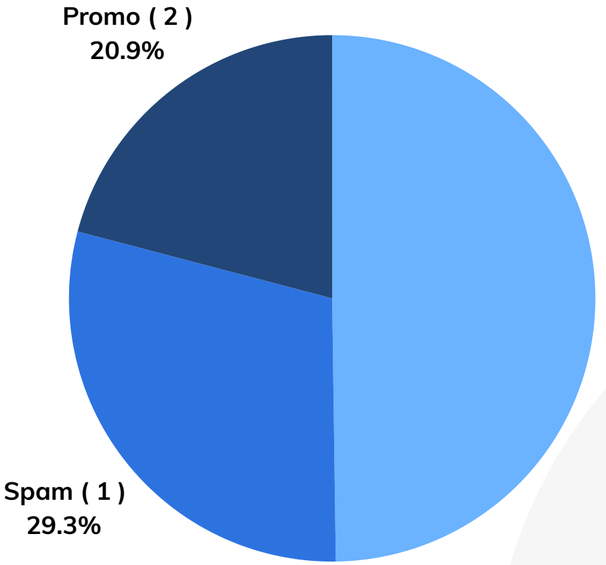
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px

#for text pre-processing
import re, string
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('wordnet')
nltk.download('stopwords')
# Find the path of the stopwords resource
stopwords_path = nltk.data.find('corpora/stopwords.zip')
print(f"The stopwords resource is located at: {stopwords_path}")
```

```
#for model-building
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB, GaussianNB
from sklearn.metrics import classification_report, f1_score, accuracy_score, confusion_matrix
from sklearn.metrics import roc_curve, auc, roc_auc_score

# bag of words
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from wordcloud import WordCloud

#for word embedding
import gensim
from gensim.models import Word2Vec #Word2Vec is mostly used for huge datasets
```

Data Understanding

Data Information

```
RangeIndex: 1143 entries, 0 to 1142
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Teks    1143 non-null   object
 1   label   1143 non-null   int64
dtypes: int64(1), object(1)
memory usage: 18.0+ KB
```



INSIGHT

1. Dataset terdiri dari **1143 row** data, dan **2 kolom**
2. Setiap kolom telah memiliki kesesuaian tipe data
3. Tidak ada missing value
4. Terdapat **1 duplikat data**, untuk selanjutnya akan dihapus saja
5. Distribusi class target **lack of balance**

```
df_train = df_train.drop_duplicates(subset='Teks', keep='first').reset_index(drop=True)
df_train['Teks'].duplicated().sum()
```

```
0
```

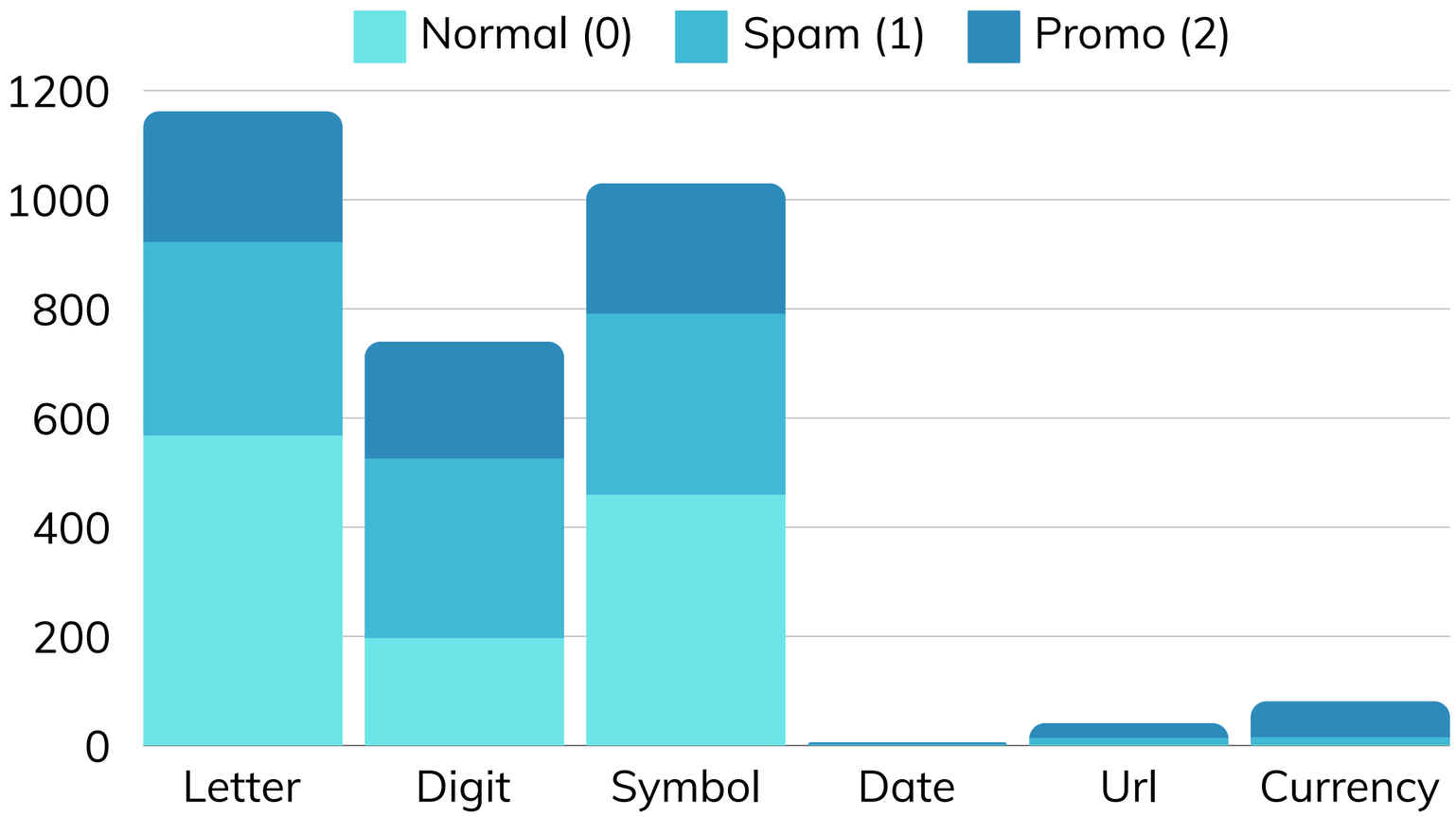
EDA Preparation

Feature Engineering

df_eda

	Teks	label	letters	digits	symbols	whitespace	url	date	currency
0	[PROMO] Beli paket Flash mulai 1GB di MY TELKO...	2	1	1	1	1	0	0	0
1	2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat A...	2	1	1	1	1	0	0	1
2	2016-07-08 11:47:11.Plg Yth, sisa kuota Flash ...	2	1	1	1	1	1	1	0
3	2016-08-07 11:29:47.Plg Yth, sisa kuota Flash ...	2	1	1	1	1	1	1	0
4	4.5GB/30 hari hanya Rp 55 Ribu Spesial buat an...	2	1	1	1	1	0	0	1

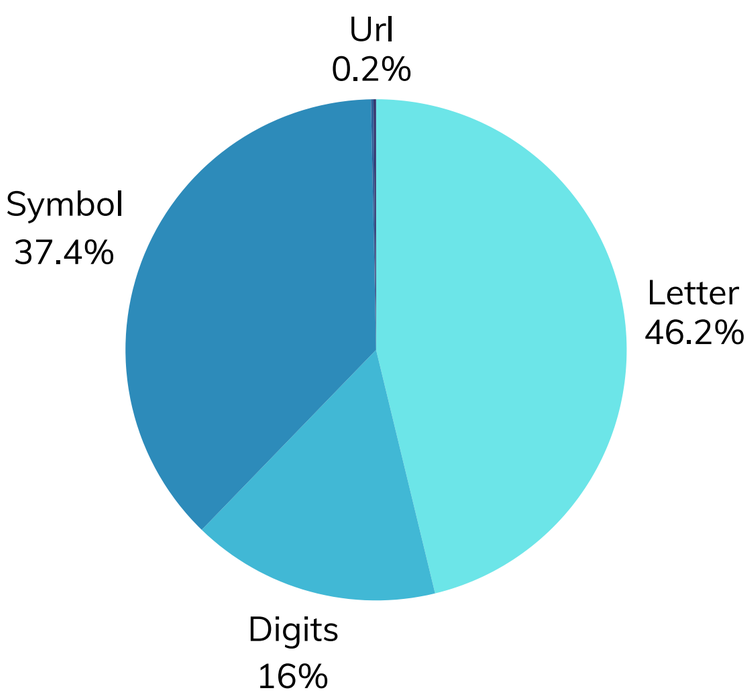
Exploratory Data Analysis



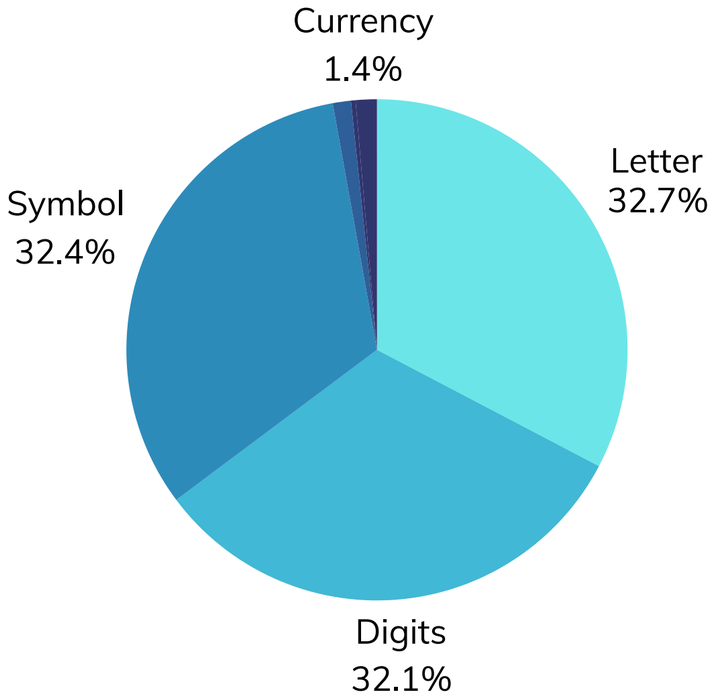
Secara umum, untuk setiap jenis pesan mengandung elemen huruf, angka, dan juga simbol, kemudian hanya sebagian kecil dari pesan spam dan promo yang mengandung elemen date, url, dan currency.

Untuk pesan normal, dari 568 data yang ada hanya 3 pesan yang mengandung url dan curency, namun tidak sama sekali mengandung elemen tanggal.

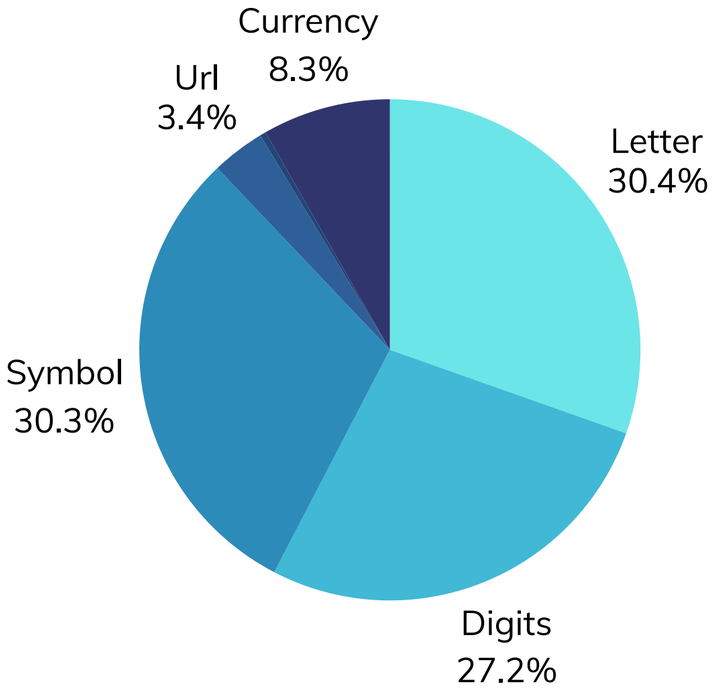
Exploratory Data Analysis



Pesan Normal

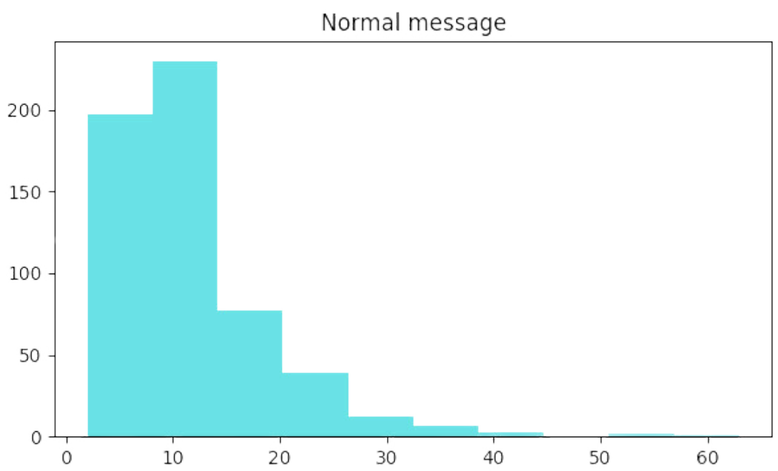


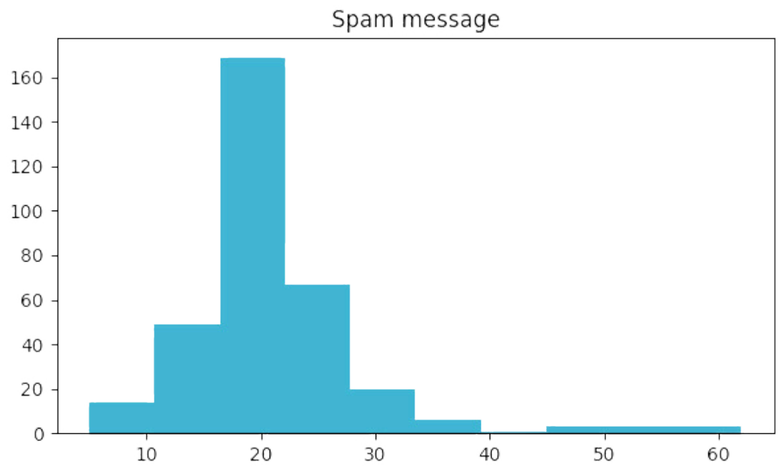
Pesan Spam

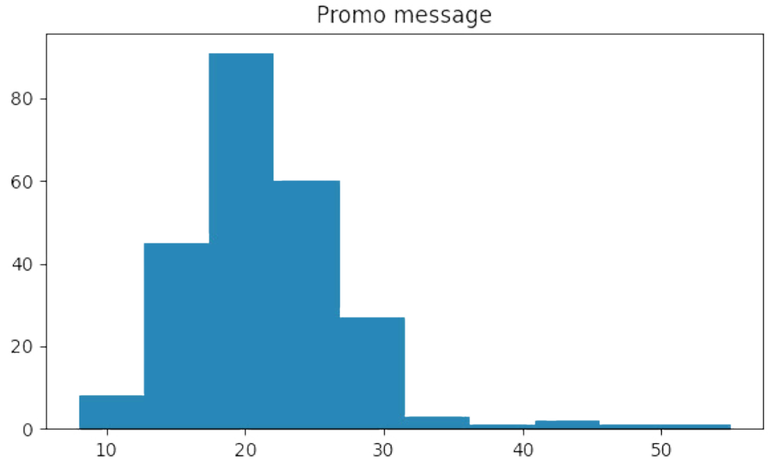


Pesan Promosi

Exploratory Data Analysis







Pesan Normal

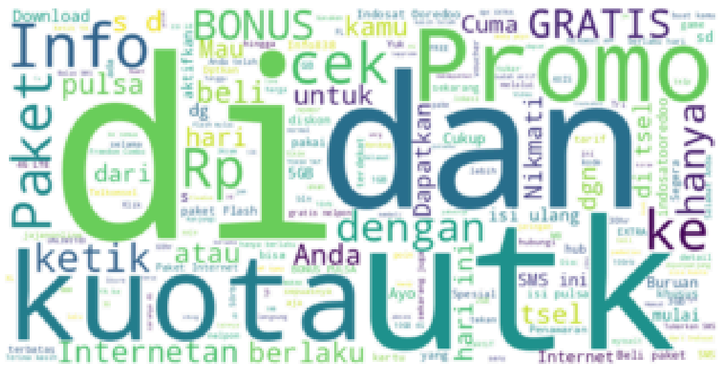
Pesan Spam

Pesan Promosi

Exploratory Data Analysis

Pesan Normal

Pesan Spam



Pesan Promosi

Data Preprocessing

	Teks
0	[PROMO] Beli paket Flash mulai 1GB di MY TELKO...
1	2.5 GB/30 hari hanya Rp 35 Ribu Spesial buat A...
2	2016-07-08 11:47:11.Plg Yth, sisa kuota Flash ...
3	2016-08-07 11:29:47.Plg Yth, sisa kuota Flash ...
4	4.5GB/30 hari hanya Rp 55 Ribu Spesial buat an...

Preprocessing 1

```python
def preprocess(text, slang_dict):
    # 1. Mengubah teks menjadi huruf kecil
    text = text.lower()

    # 2. Menghapus beberapa elemen
    text = text.strip()  # Menghapus spasi di awal/akhir teks
    text = re.compile('<.*?>').sub('', text)  # Menghapus tag HTML
    text = re.sub(r"http\S+|www\S+|https\S+", '', text, flags=re.MULTILINE) # Menghapus link
    text = re.compile('[%s]' % re.escape(string.punctuation)).sub(' ', text)  # Mengganti tanda baca dengan spasi
    text = re.sub('\s+', ' ', text)  # Menghapus spasi tambahan dan tab
    text = re.sub(r'\[[0-9]*\]', ' ', text)  # Menghapus angka dalam kurung siku
    text = re.sub(r'[^\w\s]', '', text)  # Menghapus karakter non-alfabet
    text = re.sub(r'\d', ' ', text)  # Menghapus angka
    text = re.sub(r'\s+', ' ', text)  # Menghapus spasi berlebih

    # 3. Menangani Kata Slang Bahasa Indonesia
    words = text.split()  # Split teks menjadi kata-kata
    processed_words = [slang_dict.get(word, word) for word in words]  # Ganti slang dengan kata asli
    text = ' '.join(processed_words)  # Gabungkan kembali menjadi satu teks

    return text
```

Preprocessing 2

```python
# Inisialisasi stemmer dari Sastrawi
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# Stopword Bahasa Indonesia
stop_words = set(stopwords.words('indonesian'))

# Fungsi untuk preprocessing gabungan
def preprocess_text(text):
    # Stemming & Lemmatization
    stemmed_text = stemmer.stem(text)

    # Tokenization
    tokens = stemmed_text.split()

    # Stopword Removal
    filtered_tokens = [word for word in tokens if word not in stop_words]

    # Gabungkan kembali ke teks
    processed_text = ' '.join(filtered_tokens)
    return processed_text
```

RESULT

clean_text
promo beli paket flash mulai gb di my telkomse...
gb hari hanya rp ribu spesial buat anda yang t...
paling yth sisa kuota flash anda kb download m...
paling yth sisa kuota flash anda kb download m...
gb hari hanya rp ribu spesial buat anda yang t...

clean_text
promo beli paket flash gb my telkomsel app ext...
gb rp ribu spesial pilih aktif promo nov buru ...
yth sisa kuota flash kb download mytelkomsel a...
yth sisa kuota flash kb download mytelkomsel a...
gb rp ribu spesial pilih aktif buru skb

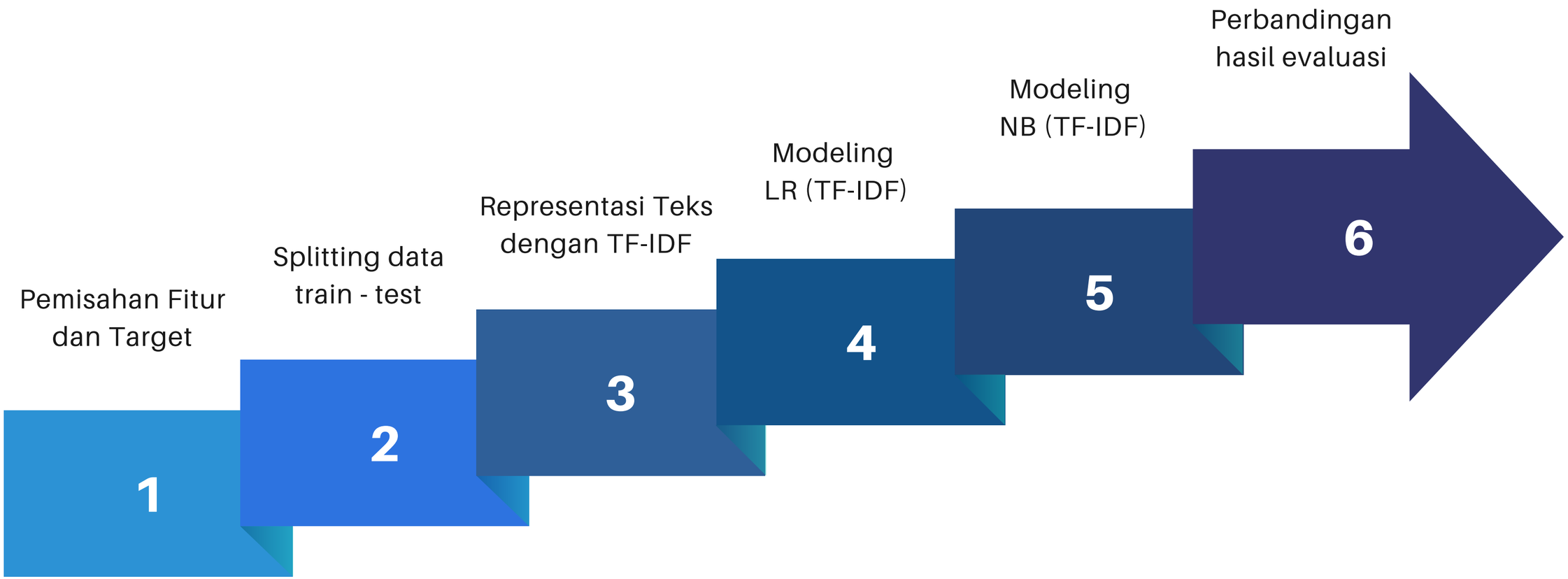
WordCloud before and after preprocessing

Before

Pesan Normal **Pesan Spam** **Pesan Promosi**

After

Modeling & Evaluation



Modeling & Evaluation

Logistic Regression (TF-IDF)

Class	Precission	Recall	F1-Score	Support
0	0.88	0.95	0.92	132
1	0.94	0.82	0.88	94
2	0.84	0.85	0.84	60
Accuracy	-	-	0.89	286
Macro Avg	0.89	0.87	0.88	286
Weighted Avg	0.89	0.89	0.89	286

Naive Bayes (TF-IDF)

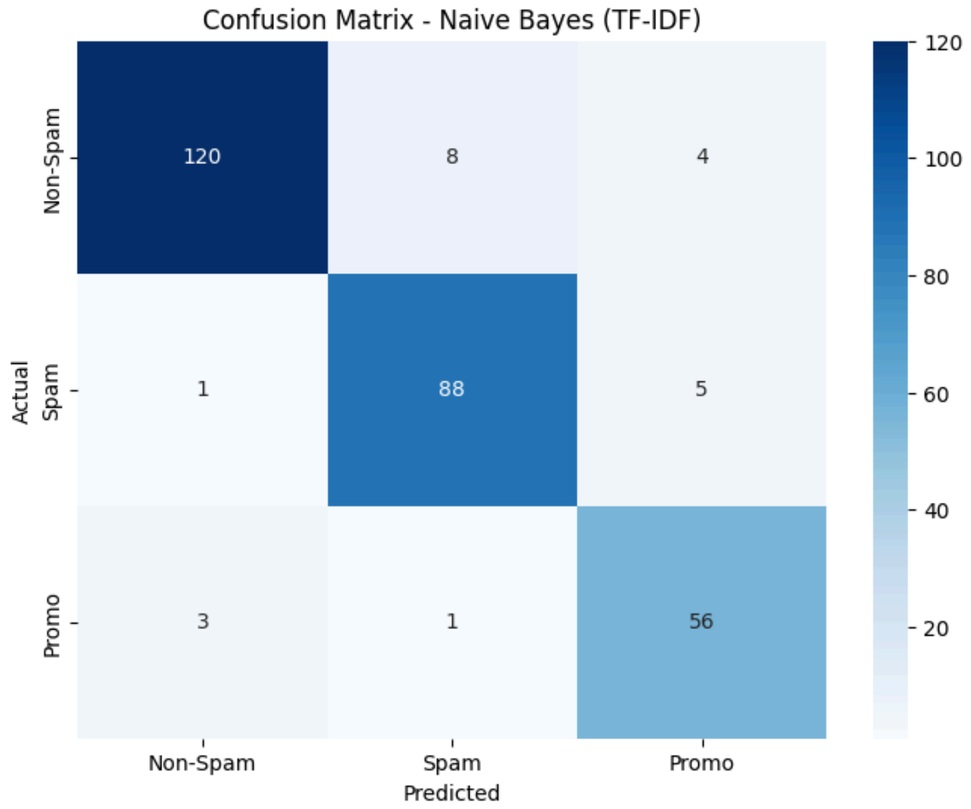
Class	Precission	Recall	F1-Score	Support
0	0.97	0.91	0.94	132
1	0.91	0.94	0.92	94
2	0.86	0.93	0.90	60
Accuracy	-	-	0.92	286
Macro Avg	0.91	0.93	0.92	286
Weighted Avg	0.93	0.92	0.92	286

Modeling & Evaluation

Rekapitulasi (Per Kelas)

Kelas	TP	FN	FP	TN
Non-Spam	120	12	4	144
Spam	88	6	9	176
Promo	56	4	9	208

Confusion Matrix



Testing Model Classification

Input data

df_new_test

	text	expected_class
0	Gratis pulsa Rp50.000 untuk pelanggan setia! K...	1
1	Rapat internal akan diadakan besok pukul 10:00...	0
2	Diskon besar-besaran hanya hari ini! Dapatkan ...	2
3	Hai, bagaimana kabarmu? Lama tak bertemu, kita...	0
4	Selamat! Anda memenangkan undian senilai Rp1.0...	1

Predict Naive Bayes (TF-IDF)

```
# Preprocessing data baru (sama seperti data training)
df_new_test['clean_text'] = df_new_test['text'].apply(lambda x: preprocess_final(x, slang_dict_id))
tfidf_new = tfidf.transform(df_new_test['clean_text'])  # Transformasikan dengan TF-IDF vectorizer yang sama

# Prediksi dengan model Naive Bayes
predictions = nb_tfidf.predict(tfidf_new)

# Menambahkan prediksi ke DataFrame
df_new_test['predicted_label'] = predictions

# Hasil
df_new_test[['text', 'predicted_label', 'expected_class']]
```

The Result

	text	predicted_label	expected_class	prob_class_0	prob_class_1	prob_class_2
0	Gratis pulsa Rp50.000 untuk pelanggan setia! K...	1	1	0.083260	0.509703	0.407038
1	Rapat internal akan diadakan besok pukul 10:00...	0	0	0.836652	0.086752	0.076596
2	Diskon besar-besaran hanya hari ini! Dapatkan ...	2	2	0.287849	0.310984	0.401167
3	Hai, bagaimana kabarmu? Lama tak bertemu, kita...	0	0	0.562454	0.237140	0.200405
4	Selamat! Anda memenangkan undian senilai Rp1.0...	1	1	0.048932	0.873959	0.077109
5	Promo spesial hari ibu! Dapatkan bunga segar h...	2	2	0.265351	0.347095	0.387554
6	Tolong kirimkan dokumen tersebut sebelum pukul...	0	0	0.488696	0.406503	0.104801
7	Update keamanan: Jangan bagikan OTP kepada sia...	0	0	0.493997	0.248316	0.257687

Thank You

Link Github

Link Google Colab