A TF-IDF Based Machine Learning Approach to Detect Emotion from Text

Kelompok 8:
Vio Albert Ferdinand - 2440017126
Francis Alexander - 2440062161
Edwin Ario Abdiwijaya - 2440062155

Background Problem

Komunikasi berbasis teks tersirat dengan beragam emosi. Seiring dengan perkembangan teknologi, kebutuhan untuk mendeteksi emosi melalui teks menjadi hal yang mendesak. Sayangnya, hal ini dapat menjadi tantangan bagi mesin untuk memahami emosi dari komunikasi tertulis. Mesin perlu berhati-hati dalam mengartikan emosi karena suatu tulisan dapat memiliki beragam makna.

Solution

1. Preprocessing
    - Data Cleaning
    - Lemmatization
    - Stop Words Removal
2. Feature Extraction
    - TF-IDF
    - N-gram
3. Classification
    - Multinomial Naive Bayes
    - Random Forest Classifier
    - Linear Support Vector Machine (SVM)
4. Performance Evaluation
    - Accuracy
    - Precision
    - Recall
    - F1-Score

Preprocessing

Data Cleaning

- Membuang data duplikat
- melakukan ekspansi contractions
- menghilangkan kata-kata tidak bermakna

Lemmatization

Melakukan perubahan bentuk kata berdasarkan Part of Speech (POS)

Stop Words Removal

Menghapus kata-kata pada kalimat yang ada di list kata-kata pada stop words.

Feature Extraction

- TF-IDF: suatu alat ukur untuk menentukan relevansi dari suatu term di dalam sekumpulan dokumen.
- N-gram: menggunakan term yang berisi n kata. Dalam solusi ini, digunakan n = 1 dan n = 2 (unigram dan bigram)

Classification

Multinomial Naïve Bayes

Mengklasifikasikan sebuah fitur dengan cara memilih kelas yang memiliki probabilitas posterior tertinggi

Random Forest Classifier

Mengklasifikasikan dengan cara membuat banyak decision tree berdasarkan sekumpulan vektor fitur yang dipilih secara random

Linear Support Vector Machine

Sebuah klasifikasi yang bergantung pada vektor fitur untuk membuat sebuah garis yang optimal dalam hyperplane yang membagi menjadi kelas-kelas yang berbeda

Performance Evaluation

Accuracy

$$\frac{True\ Positive + True\ Negative}{Number\ of\ Samples}$$

Precision

$$\frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall

$$\frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1-Score

$$2 \times \frac{Precision * Recall}{Precision + Recall}$$

Implementation Details

Preparation & Preprocessing

Data Retrieval

Dataset: emotion dataset retrieved by HuggingFace using CARER
Library used: numpy, pandas

Data Cleaning

Library used: contractions, Python Standard Library

Lemmatization

WordNetLemmatizer()
Library used: NLTK

Stop Words Removal

NLTK english stop words excluding negation words
Library used: NLTK

Implementation Details

Feature Extraction

TF-IDF + N-gram

Function: TfidfVectorizer()
Parameter:

- ngram_range: (1,2)
- min_df: 10
- norm: ‘l2’
- analyzer: ‘word'
- sublinear_tf: True

Library used: sklearn

Implementation Details

Machine Learning Models

Multinomal Naive Bayes

Function: MultinomialNB()
Parameters: default
Library used: sklearn

Random Forest Classifier

Function: RandomForestClassifier()
Parameters: default
Library used: sklearn

Linear SVM

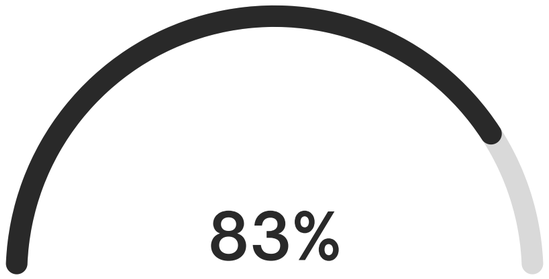
Function: RandomForestClassifier()
Parameters:

- loss: squared_hinge
- dual: True
- tol: 1e-4
- C: 0.5
- class_weight: None
- max_iter: 1000

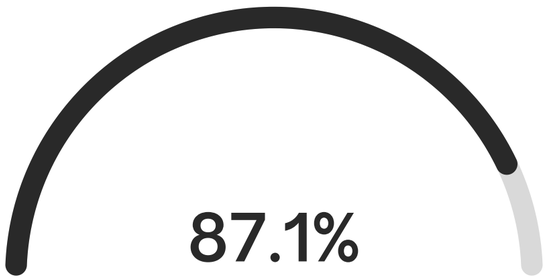
Library used: sklearn

Accuracy of Each Model

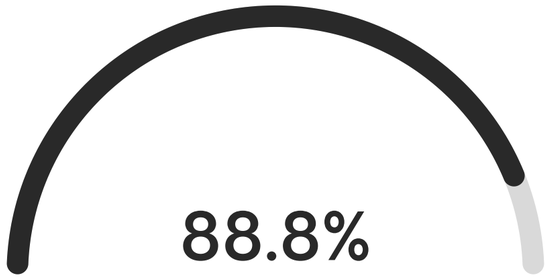
Multinomial Naive Bayes



Random Forest Classifier



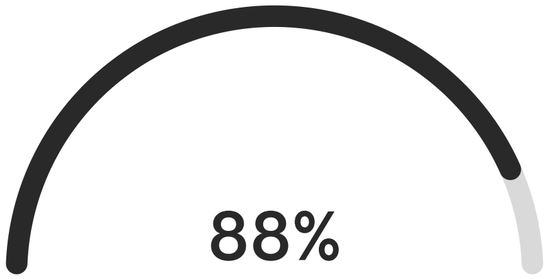
Linear SVM



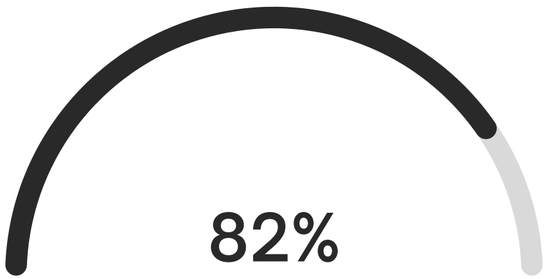
Linear SVM memiliki akurasi tertinggi dari ketiga model yang diujikan

Precision of Each Model

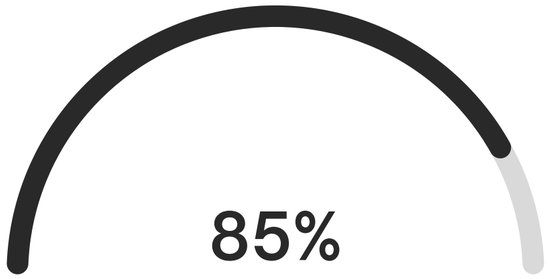
Multinomial Naive Bayes



Random Forest Classifier



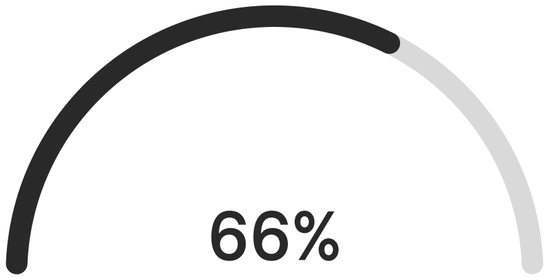
Linear SVM



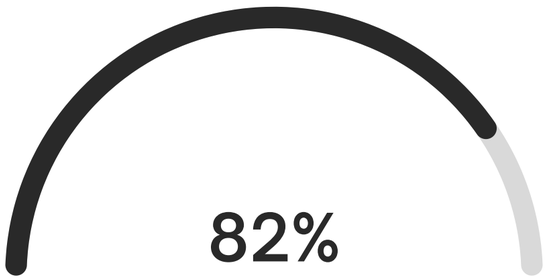
Multinomal Naive Bayes memiliki precision tertinggi dari ketiga model yang diujikan

Recall of Each Model

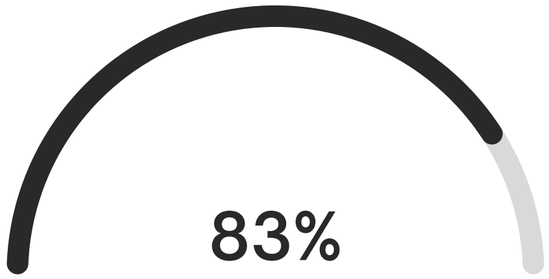
Multinomial Naive Bayes



Random Forest Classifier



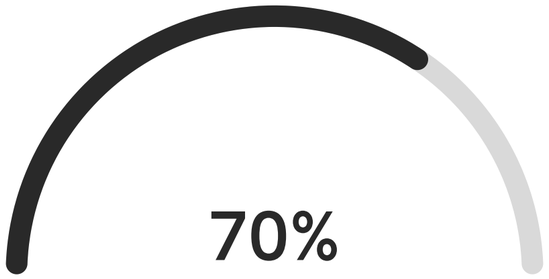
Linear SVM



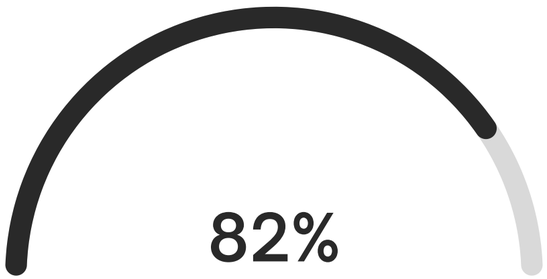
Linear SVM memiliki recall tertinggi dari ketiga model yang diujikan

F1-score of Each Model

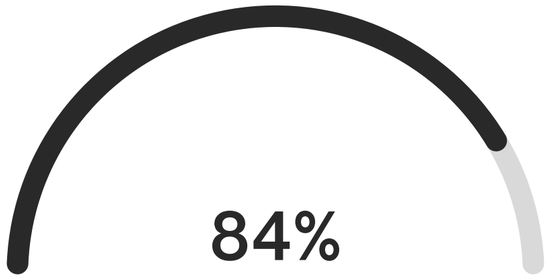
Multinomial Naive Bayes



Random Forest Classifier



Linear SVM



Linear SVM memiliki F1-score tertinggi dari ketiga model yang diujikan

Linear SVM (tuned)

Accuracy

Precision

Recall

F1-Score

Demo

https://share.streamlit.io/vioalbert/emotion-detection/app.py


Thank
You