# Report Classify Candidate Pairs of Acronyms and Expansions

**Nama:** Najwan Yusnianda
**NIM:** 2408207010029

## 1. Introduction

Selama beberapa dekade sebelumnya, identifikasi cerdas pasangan akronim dan perluasan dari korpus besar telah menarik perhatian penelitian yang cukup besar, terutama di bidang penambangan teks,ekstraksi entitas, dan pencarian informasi . Salah satu penelitian oleh Taufik et al. **[1]** memperkenalkan delapan fitur vektor untuk menggambarkan pasangan akronim dan ekspansinya. Dengan fitur tersebut, Machine Learning dapat mencapai akurasi tinggi dalam klasifikasi pasangan akronim dan kepanjangannya. Penelitian kali ini bertujuan untuk menentukan metode klasifikasi supervised learning terbaik dengan fitur-fitur tersebut serta membandingkannya dengan metode klasifikasi deep learning berbasis transformer menggunakan Bidirectional Encoder Representations from Transformers (BERT)

## 2. Methodology

### 2.1. Dataset

Dataset yang digunakan dalam penelitian ini berasal dari dataacro [1] yang mempunyai 8 fitur hasi dari ekstraksi fitur pasangan akronim yang telah dikumpulak. Dataset ini terdiri dari training set (4000 sampel ) dengan testing set ( 1099 sampel). berikut adalah beberapa contoh training set:

```
BUMD=>Usaha Milik -1 1:0.91829583405449 2:1 3:-0.666666666666667 4:0 5:1 6:0.5 7:0
8:0.393089881055403
TNI=>meminjam senjata dari oknum -1 1:1 2:0.5 3:-2 4:0 5:0.75 6:0 7:0
8:0.0357142857142857
PKI=>Panitia Pengawas -1 1:0.970950594454669 2:1 3:-1 4:0.5 5:1
6:0.333333333333333 7:0 8:0.400611989684
MA=>putusan Mahkamah -1 1:1 2:0.75 3:-2 4:0 5:1 6:1 7:1 8:0.392857142857143
```

Data tersebut yang dikumpulkan selanjutnya dilakukan preprocessing untuk memisahkan antara fitur dan label dan text agar selanjutnya dapat digunakan untuk supervised learning. label yang sebelumnya terdiri dari -1 (negative class) dan 1 (positive class) diubah menjadi (0,1) agar memudahkan saat pretraining

| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | label |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.918296 | 1 | -0.666667 | 0 | 1 | 0.5 | 0 | 0.39309 | 0 |
| 1 | 1 | 0.5 | -2 | 0 | 0.75 | 0 | 0 | 0.0357143 | 0 |
| 2 | 0.970951 | 1 | -1 | 0.5 | 1 | 0.333333 | 0 | 0.400612 | 0 |
| 3 | 1 | 0.75 | -2 | 0 | 1 | 1 | 1 | 0.392857 | 0 |

| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | label |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 0.970951 | 0.666667 | -2.5 | 0 | 1 | 0 | 0 | 0.0196596 | 0 |

*Table 1: Dataset yang digunakan untuk supervised learning*

| | Fitur teks | label |
|---|---|---|
| 0 | BUMD=>Usaha Milik | 0 |
| 1 | TNI=>meminjam senjata dari oknum | 0 |
| 2 | PKI=>Panitia Pengawas | 0 |
| 3 | MA=>putusan Mahkamah | 0 |
| 4 | TI=>com Mati body | 0 |

*Table 2: Dataset yang digunakan untuk klasifikasi teks dengan BERT*

Rincian fitur - fitur tersebut adalah sebafai berikut:

- **Fitur 1** : Korelasi antara jumlah total karakter dalam akronim dan total jumlah kata dalam ekspansi
- **Fitur 2** : Jumlah kata dalam ekspansi yang menggunakan huruf besar pada awal kata
- **Fitur 3** : Penimbang kecocokan huruf-huruf dalam akronim dan ekspansi/kepanjangannya, tidak termasuk kata sambung
- **Fitur 4** : Penimbang korelasi antara huruf pertama dan terakhir dari akronim.
- **Fitur 5** : Nilai Penalti kepada akronim yang mengandung banyak preposisi (kata depan) dan konjungsi (kata penghubung)
- **Fitur 6** : Rasio kecocokan yang tepat antara karakter dalam ekspansi dan karakter dalam akronim
- **Fitur 7** : Nilai Pembeda antara rasio kecocokan yang akurat (Fitur 6) dan rasio yang tidak akurat
- **Fitur 8** : rata-rata dari Fitur 1 hingga Fitur 7.

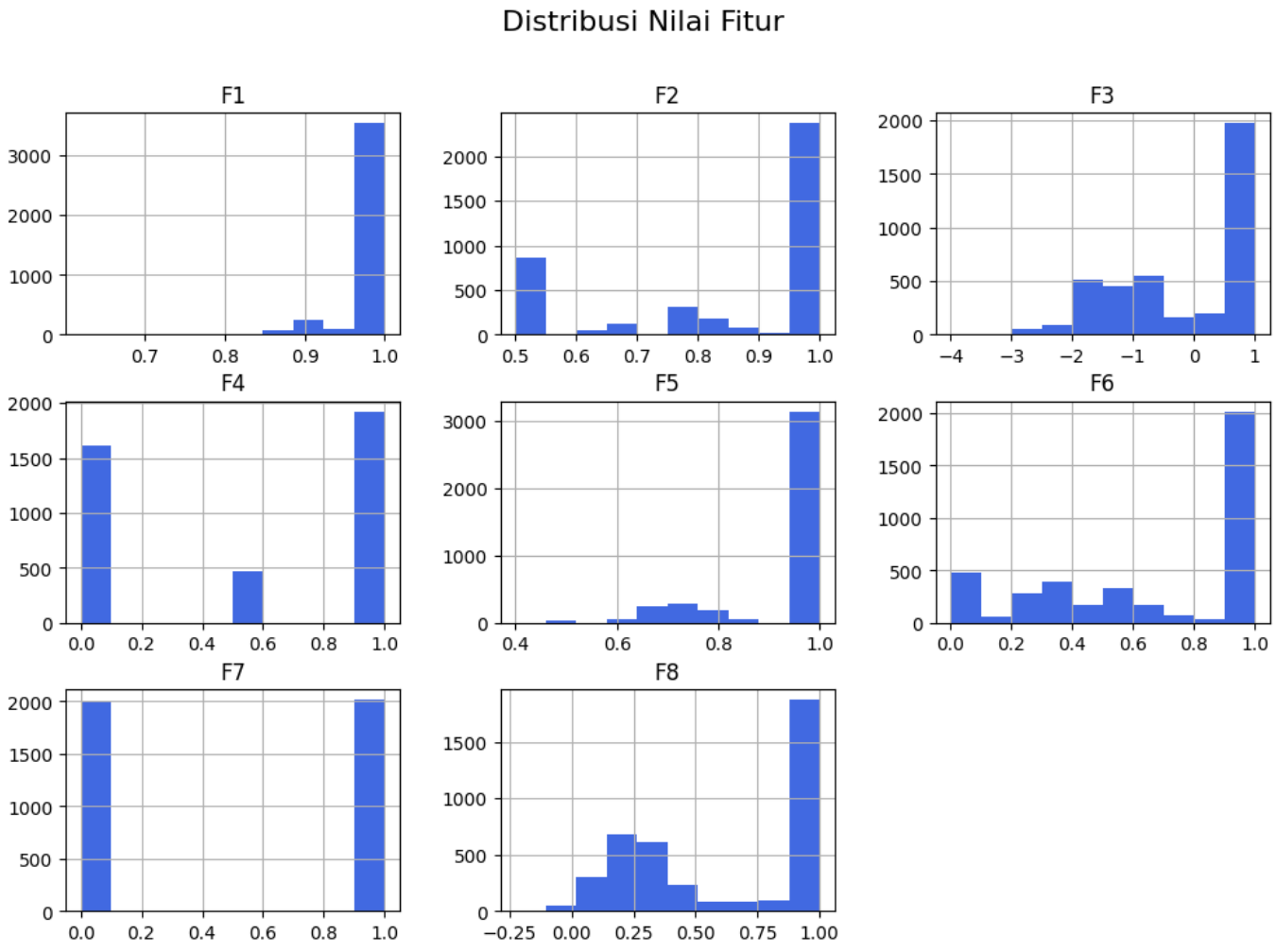
## 2.2. Data Description

Sebelum melakukan pemodelan, penting untuk memahami karakteristik data melalui analisis statistik deskriptif. Berikut adalah analisis deskriptif dari data (training set) yang digunakan:

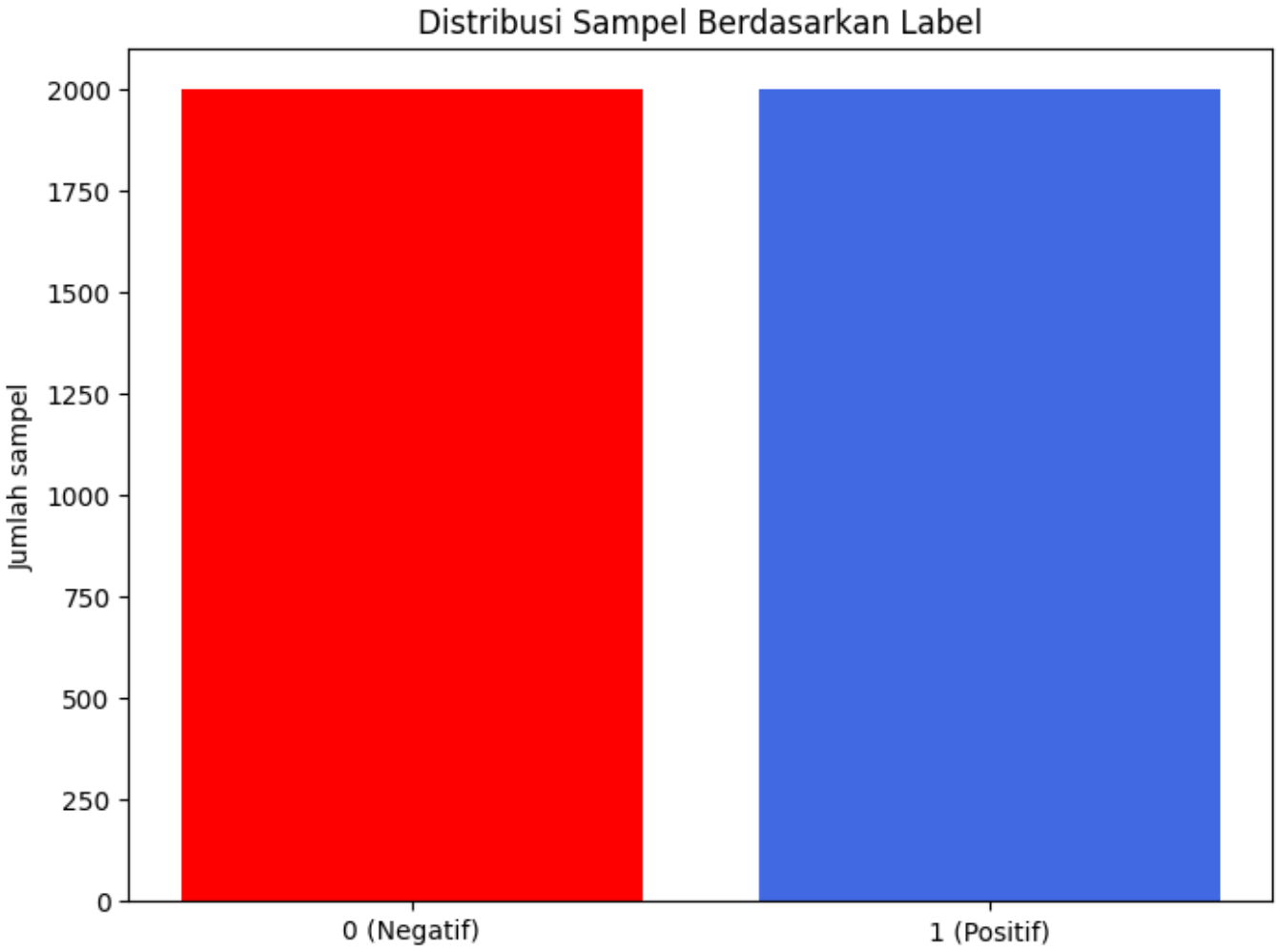
```
|       |           F1 |          F2 |          F3 |          F4 |          F5 |
F6 |          F7 |          F8 |       label |
|:------|-------------:|------------:|------------:|------------:|------------:|--
----------:|------------:|------------:|------------:|
| count | 4000         | 4000        | 4000        | 4000        | 4000        |
4000        | 4000        | 4000        | 4000        |
| mean  |    0.983223  |    0.847115 |   -0.147867 |    0.538875 |    0.939538 |
0.664874 |    0.503    |    0.618394 |    0.5      |
| std   |    0.0331568 |    0.205513 |    1.21748  |    0.468088 |    0.120776 |
0.375917 |    0.500054 |    0.366575 |    0.500063 |
| min   |    0.619382  |    0.5      |   -4        |    0        |    0.4      |
0        |    0        |   -0.225958 |    0        |
| 25%   |    0.970951  |    0.666667 |   -1.28571  |    0        |    1        |
0.333333 |    0        |    0.257755 |    0        |
```

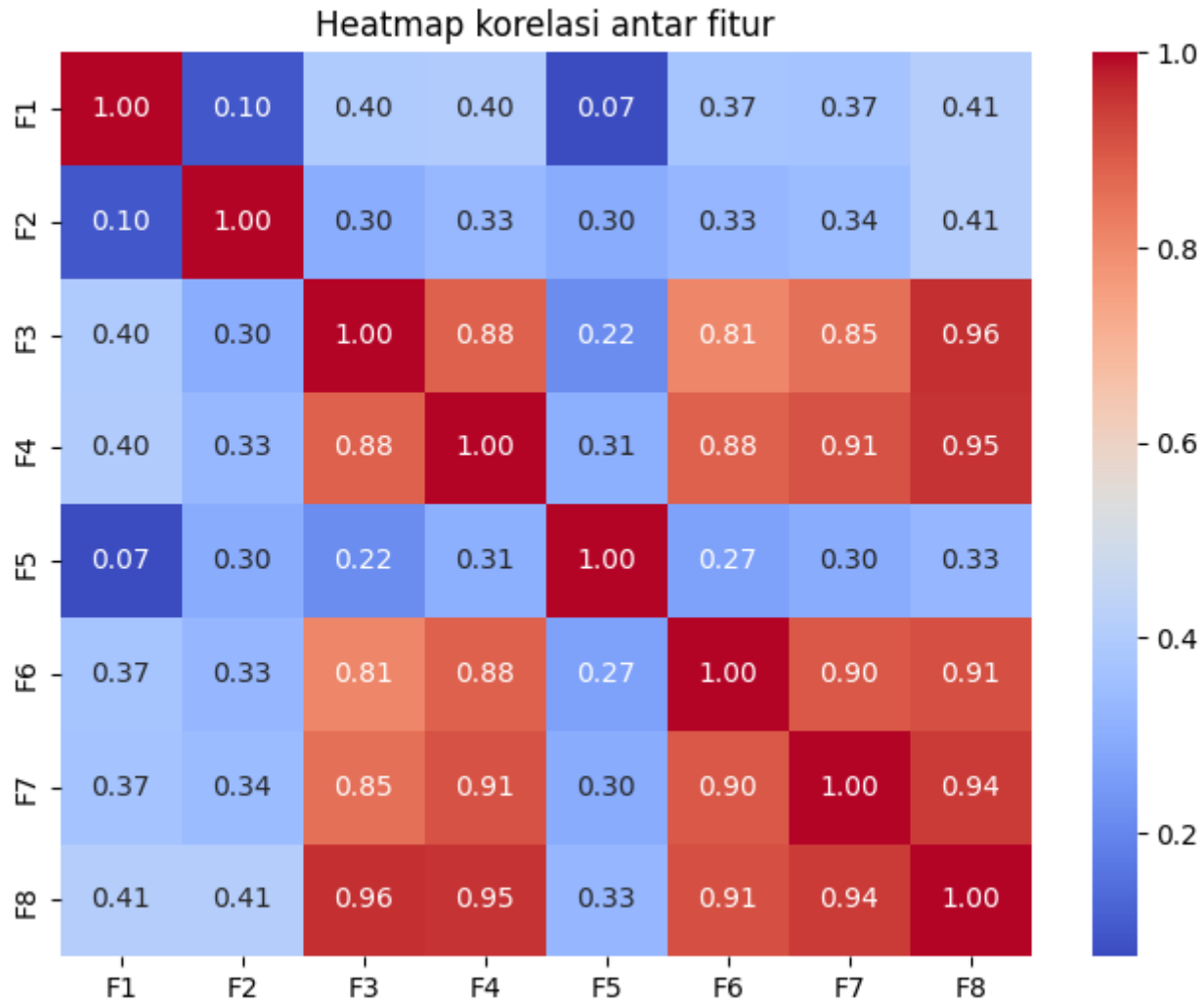
```
| 50%   |    1          |    1          |    0.333333 |    0.5        |    1          |
1          |    1          |    0.713231 |    0.5        |
| 75%   |    1          |    1          |    1          |    1          |    1          |
1          |    1          |    1          |    1          |
| max   |    1          |    1          |    1          |    1          |    1          |
1          |    1          |    1          |    1          |
```

1. **Korelasi antara fitur** : Untuk mengetahui hubungan ant




Berdasarkan gambar tersebut ....


Distribusi Sampel Berdasarkan Label
Jumlah sampel
2000
1750
1500
1250
1000
750
500
250
0
0 (Negatif)
1 (Positif)




## 2.3. Model and Algorithms

Penelitian ini menggunakan model dan algoritma supervised learning dengan menggunakan delapan fitur tersebut untuk menemukan model terbaik kemudian dibandingkan dengan model berbasis Transformer yaitu BERT (Bidirectional Encoder Representations from Transformers). Model supervised learning yang digunakan adalah sebagai berikut:

1. **Support Vector Machine**
2. **K-Nearest Neighbor (KNN)**
3. **Naive Bayes**
4. **Decision Tree**

Selanjutnya, model berbasis BERT ) digunakan sebagai pembanding. Model ini memanfaatkan arsitektur transformer untuk menangkap konteks dari kata secara lebih mendalam.

## 2.4. Experimental Setup

Eksperimen dilakukan menggunakan Python dengan framework Scikit-Learn untuk supervised learning dan PyTorch untuk fine-tuning BERT.
Tahapan Supervised Learning adalah sebagai berikut:

- Penyiapan Dataset: Dataset yang digunakan merupakan data akronim dan ekspansinya yang telah dilakukan ekstraksi fitur menjadi delapan fitur penting (F1 s.d F8).

- Tuning Hyperparameter: dilakukan dengan GridSearchCV untuk menentukan hyperparameter terbaik.
- Model dilatih dengan 4000 sampel training set dan 1099 sampel testing set.
- Selanjutnya setiap model dilakukan evaluasi menggunakan metrik evaluasi yang sudah ditentukan.

Tahapan untuk Deep Learning berbasis transformer dengan BERT adalah sebagai berikut:

- Penyiapan Dataset: Dataset yang digunakan terdiri dari fitur 'akronim=>ekspansi' serta label yang perlu dilakukan preprocessing menggunakan tokenizer dari BERT dan menyimpannya dalam dataset custom dalam bentuk tensor agar bisa dilatih dalam pytorch
- Inisiasi model : Mengimpor model BERT (pre-trained) sebelumnya dari Hugging Face
- Menyiapkan trainer dan fine tuning : Model pretrained BERT dilatih dengan 4000 sampel training set
- Selanjutnya model dilakukan evaluasi menggunakan metrik evaluasi yang sudah ditentukan.

### 2.5. Evaluation Metrics

Tahapan evalusasi model, metric yang digunakan adalah sebagai berikut:

- **F1-score**: Untuk menilai keseimbangan antara presisi dan recall.
- **Accuracy**: Untuk melihat persentase prediksi yang benar.
- **Confusion Matrix**: Untuk analisis kesalahan klasifikasi.

## 3. Results and Discussion

Berikut adalah hasil penelitian:

## 4. Conclusion

Hasilnya menunjukkan bahwa...

## 5. References

[1] Abidin TF, Mahazir A, Subianto M, Munadi K, Ferdhiana R. Recognizing Indonesian Acronym and Expansion Pairs with Supervised Learning and MapReduce. Information. 2020; 11(4):210. https://doi.org/10.3390/info11040210