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

	Fitur 1	Fitur 2	Fitur 3	Fitur 4	Fitur 5	Fitur 6	Fitur 7	Fitur 8	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

	Fitur 1	Fitur 2	Fitur 3	Fitur 4	Fitur 5	Fitur 6	Fitur 7	Fitur 8	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

Adapun penjelasan 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. dan kata depan

2.2. Data Description

Kami menggunakan delapan fitur vektor yang diperkenalkan oleh [1] untuk mewakili pasangan akronim dan ekspansi.

2.3. Model and Algorithms

Kami membandingkan beberapa model supervised learning, yaitu:

- **SVM** dengan kernel linear dan RBF
- **Random Forest** dengan jumlah pohon berbeda
- **BERT** sebagai pendekatan berbasis deep learning

2.4. Experimental Setup

Eksperimen dilakukan menggunakan Python dengan framework Scikit-Learn dan PyTorch.

- Hyperparameter tuning dilakukan dengan GridSearchCV.
- Model dilatih dengan 10-fold cross-validation.

- Training dilakukan pada **NVIDIA GTX 1630 GPU** dengan batch size 32.

2.5. Evaluation Metrics

Untuk mengevaluasi performa model, kami menggunakan metrik 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:

 Grafik

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