KLASIFIKASI BAKU LAPANGAN USAHA INDONESIA(KBLI) MENGGUNAKAN *NATURAL LANGUAGE PROCESSING*

Lidya Elisabet Theogracia Silitonga[1], I Putu Gede Hendra Suputra[2], dan I Gede Santi Astawa[3]

ABSTRAK

Klasifikasi Baku Lapangan Usaha Indonesia(KBLI) merupakan klasifikasi baku kegiatan ekonomi yang terdapat di Indonesia. KBLI disusun untuk menyediakan satu set kerangka klasifikasi kegiatan ekonomi yang komprehensif di Indonesia agar dapat digunakan untuk penyeragaman, pengumpulan, pengolahan, penyajian, dan analisis data statistik menurut kegiatan ekonomi serta untuk mempelajari keadaan atau perilaku ekonomi menurut kegiatan ekonomi. Badan Pusat Statistik Provinsi Bali memiliki data baku lapangan usaha di Indonesia, akan tetapi data tersebut masih berbentuk data mentah dan belum diklasifikasikan secara menyeluruh. Implementasikan *Natural Language Processing* dalam proses klasifikasi data baku lapangan usaha Indonesia yang bertujuan untuk meningkatkan akurasi model dan meningkatkan performa klasifikasi yang akurat.

Kata kunci : klasifikasi baku lapangan usaha Indonesia

ABSTRACT

The Standard Classification of Business Fields in Indonesia (KBLI) is a standard classification of economic activities in Indonesia. The KBLI was developed to provide a comprehensive set of economic activity classification frameworks in Indonesia so that they can be used for standardizing, collecting, processing, presenting, and analyzing statistical data according to economic activity and for studying economic conditions or behavior according to economic activity. The Central Bureau of Statistics for the Province of Bali has standard data on business fields in Indonesia, but this data is still in the form of raw data and has not been thoroughly classified. Implement transfer learning in the process of classifying raw data for business fields in Indonesia which aims to improve model accuracy and improve accurate classification performance.

Keywords: Indonesian Standard Industrial Classification

1. PENDAHULUAN

Klasifikasi Baku Lapangan Usaha di Indonesia(KBLI) merupakan klasifikasi baku kegiatan ekonomi yang terdapat di Indonesia. KBLI disusun untuk menyediakan satu set kerangka klasifikasi kegiatan ekonomi yang komprehensif di Indonesia agar dapat digunakan untuk penyeragaman, pengumpulan, pengolahan, penyajian, dan analisis data statistik menurut kegiatan ekonomi serta untuk mempelajari keadaan atau perilaku ekonomi menurut kegiatan ekonomi. Badan Pusat Statistik Provinsi Bali memiliki data baku lapangan usaha di Indonesia, akan tetapi data tersebut masih berbentuk data mentah dan belum diklasifikasikan secara menyeluruh. Oleh karena itu, penulis mengimplementasikan *Natural Language Processing*(NLP) dalam

[1] *Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana*
[2] *Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana*
[3] *Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana*
Submitted: 8 Oktober 2023 *Revised: 2 November 2023* *Accepted: 3 November 2023*

proses klasifikasi data baku lapangan usaha di Indonesia yang bertujuan untuk meningkatkan akurasi model dan meningkatkan performa klasifikasi yang akurat.

2. METODE PELAKSANAAN

Adapun tahapan proses klasifikasi baku lapangan usaha di Indonesia menggunakan *Natural Language Processing*, yaitu

- Pemahaman Dataset (*Understanding Dataset*)
  Pemahaman dataset yang dimiliki yang bertujuan untuk memperoleh informasi dari data tersebut.
- *Loading the Data*
  Sebelum melakukan *Loading the Data*, kita harus menginstall terlebih dahulu modul dan mengimport *libraries* yang dibutuhkan. Setelah itu, melakukan *loaded dataset* yang mana data tersebut akan disimpan setekah dibaca dari file CSV menggunakan DataFrame(df).
- Text processing
  Pada tahap ini, data tersebut akan diubah menjadi vektor dengan menggunakan TFIDF(Term Frequency-Inverse Document Frequency).
- Mengeksplor Model Klasifikasi Teks
  Adapun model klasifikasi yang digunkan, yaitu Random Forest, Linear Support Vector Machine, Multinomial Naïve Bayes, dan Logistic Regression)
- Membandingkan Performance Model Klasifikasi
  Pada tahap ini akan membandingkan ‘Mean Accuracy’ dan ‘Standar Deviation’ dari ke empat algoritma tersebut (model yang cocok untuk proses klasifikasi data tersebut).
- Evaluasi Model Klasifikasi Teks
  Pada tahap ini akan dilakukan *train* terhadap model data yang dipilih sehingga kita dapat mengevaluasi dan memeriksa *performance* data yang tidak terlihat.
- Prediksi
  Melakukan prediksi pada data yang tidak terlihat dan mengecek *performance* dari model.

3. HASIL DAN PEMBAHASAN

Berikut ini proses klasifikasi data baku lapangan usaha Indonesia menggunakan *Natural Language Processing*:

- *Loading the Data*
  Pada tahap *loading the Data*, download terlebih dahulu data yang kita miliki dengan melakukan import dataset dari google drive, setelah itu install modul yang dibutuhkan dan kita juga dapat mengecek jumlah baris dan kolom dari dataset, seperti pada gambar 3.1. Pada gambar 3.2 dibuat data frame baru yang mana pada awalnya dataset tersebut memiliki 65535 baris dan kolom, lalu pada saat membuat data frame baru terdapat dua kolom yang dipilih, yaitu B1R15A_KOD(Kode Usaha) dan B1R15A(Jenis Usaha) untuk data frame baru serta menghapus *missing value*(NaN) yang terdapat pada baris dataset.

Gambar 3.1. Install modul dan libraries & loaded dataset

Gambar 3.2. Membuat data frame baru

Setelah membuat data frame baru dan menghapus *missing value*(NaN), pada gambar 3.3 dilakukan pengecekkan kode apa saja yang terdapat di B1R15A_KOD(Kode Usaha). Lalu pada gambar 3.4, untuk mempermudah proses *training*, maka akan dilakukan beberapa perubahan pada nama kategori, yaitu B1R15A(Kode Usaha).

Gambar 3.3. Melakukan pengecekkan nilai unik

Gambar 3.4. Melakukan perubahan nama Kategori

Pada gambar 3.5 dilakukan pemetaan pada masing-masing kategori ke sebuah angka sehingga model dapat memahami dengan baik dan akan disimpan ke dalam kolom baru dengan nama 'category_id' yang mana masing-masing dari kategori direpresentasikan secara numerik. Pada gambar 3.6 dilakukan visualisasi dari data tersebut untuk melihat jumlah Jenis Usaha(B1R15A) yang terdapat pada tiap kategori(B1R15A_KOD).

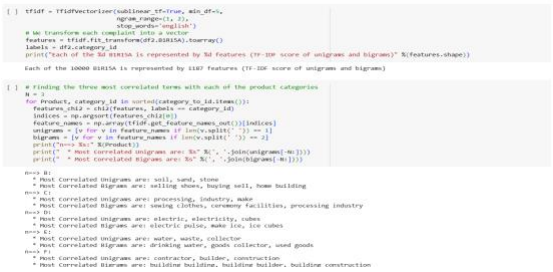
	B1R15A_KOD	B1R15A	category_id
12284	G	SELLING SHOES	0
56925	G	SELLING FOODS AND TOBACCO	0
43007	G	SELLING COURSE BOOKS	0
35330	G	SELLING SIDE FOOD	0
1131	I	SELLING FOOD IN SHOPS	1

Gambar 3.5. Map masing-masing kategori

Gambar 3.6. Visualisasi data

- *Text processing*

Pada tahap *text processing* dan pada gambar 3.7 teks diubah menjadi vektor menggunakan Term Frequency-Inverse Document Frequency (TFIDF) yang mana *TFIDFVectorizer* digunakan dengan tiga parameter, yaitu min_df digunakan untuk menghapus kata-kata yang muncul dalam jumlah file kuran dari 'min_df', sublinear_tf digunakan untuk menskalakan frekuensi dalam skala logaritmik, dan stop_words untuk menghapus kata-kata yang tidak memiliki makna.



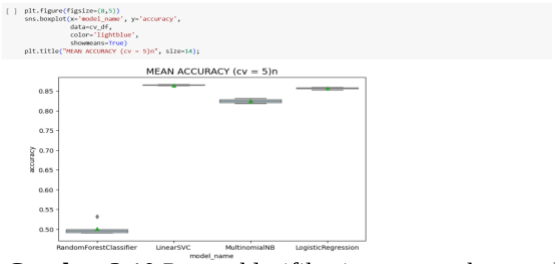
Gambar 3.7. Mengubah text menjadi vektor menggunakan TFIDF dan Menemukan korelasi

- *Exploring Multi-Classification Models*
  Pada gambar 3.8 data akan dibagi menjadi train dan test set yang mana data yang digunakan untuk training adalah 75% dan data yang digunakan untuk testing adalah 25%. Pada gambar 3.9, untuk membuat model klasifikasi terdapat beberapa algoritma yang digunakan, yaitu RandomForestClassifier, LinearSVC, MultinomialNB, dan LogisticRegression.

Gambar 3.8. Split data ke train dan test sets

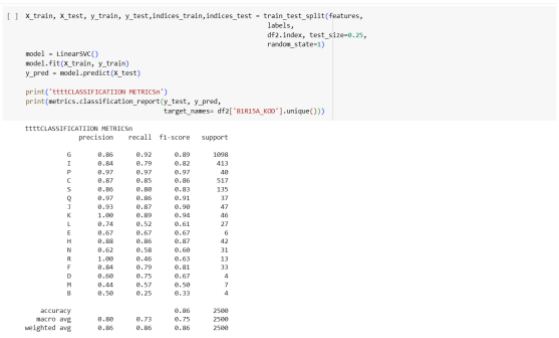
Gambar 3.9. Menghitung dan Membandingkan performance model

- *Compare Text Classification Model Performance*
  Setelah melakukan percobaan terhadap empat algoritma tersebut, akurasi dari algoritma LinearSVC lebih tinggi dari algoritma lainnya yang mana algoritma tersebut digunakan untuk model klasifikasi.



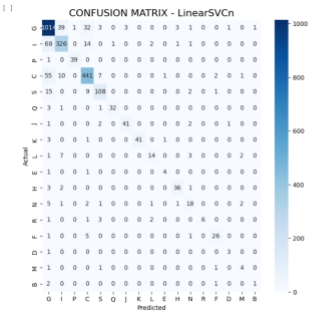
Gambar 3.10 Proses klasifikasi menggunakan model LinearSVC akurat.

- *Evaluation of Text Classification Model*
  Pada gambar 3.11 model akan dilatih menggunakan 'LinearSVC' sehingga kita dapat mengevaluasi dan mengecek performance pada data yang tidak terlihat. Lalu pada gambar 3.12 akan dilakukan plot matriks untuk memerika *miss classified prediction*. Dan gambar 3.13 merupakan hasil dari plot *confusion matrix*.



Gambar 3.11. Train Model Menggunakan Linear Support Vector Machine

Gambar 3.12. Plot confusion matrix



Gambar 3.13. Confusion Matrix

- *Prediction*
  Pada gambar 3.14 dibuat prediksi pada data yang tidak terlihat dan mengecek performance dari model yang telah dibuat.

```
[ ]  X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                         test_size=0.25,
                                                         random_state = 0)
     tfidf = TfidfVectorizer(sublinear_tf=True, min_df=5,
                             ngram_range=(1, 2),
                             stop_words='english')
     fitted_vectorizer = tfidf.fit(X_train)
     tfidf_vectorizer_vectors = fitted_vectorizer.transform(X_train)
     model = LinearSVC().fit(tfidf_vectorizer_vectors, y_train)
```

Gambar 3.14. Membuat prediksi dan Mengecek performance model

Pada gambar 3.15, misal kita memasukkan jenis usaha(B1R15A) adalah “SELLING COFFEE” maka setelah *code* di *run*, maka akan ditampilkan kode usaha(B1R15A_KOD) adalah ‘I’.

```
[ ]  B1R15A = "SELLING COFFEE"
     print(model.predict(fitted_vectorizer.transform([B1R15A])))

     ['I']
```

Gambar 3.15. Hasil Proses Klasifikasi

4. KESIMPULAN

Badan Pusat Statistik Provinsi Bali memiliki data baku lapangan usaha di Indonesia, akan tetapi data tersebut masih berbentuk data mentah dan belum diklasifikasikan secara menyeluruh. Implementasikan transfer learning dalam proses klasifikasi data baku lapangan usaha di Indonesia yang bertujuan untuk meningkatkan akurasi model, meningkatkan performa klasifikasi yang akurat.

DAFTAR PUSTAKA

Muthia, D. A. (2018). Komparasi Algoritma Klasifikasi Text Mining Untuk Analisis Sentimen Pada Review Restoran. *Jurnal PILAR Nusa Mandiri*, *14*(1), 69-74.

Halaman ini sengaja dikosongkan