POLITEKNIK STATISTIKA STIS

For Better Official Statistics

Automatic Text Categorization to Standard Classification of Indonesian Business Fields (KBLI) 2020

Jakarta, 22 September 2023

Presentation Outline

1. Introduction

Classification systems play a fundamental role in official statistics by providing structure and consistency to data collection, analysis, and reporting. They are essential tools for policymakers, researchers, and the public to understand and make informed decisions based on accurate and comparable statistical information

KBLI is a system used in Indonesia to classify and categorize economic activities and businesses based on their primary functions and outputs.

Automatic Text Categorization to Standard Classification of Indonesian Business Fields (KBLI) 2020

Lya Hulliyyatus Suadaa
Data Science Research Unit
Politeknik Statistika STIS
Jakarta, Indonesia
lya@stis.ac.id

Farid Ridho
Data Science Research Unit
Politeknik Statistika STIS
Jakarta, Indonesia
faridr@stis.ac.id

Anugerah Karta Monika
Data Science Research Unit
Politeknik Statistika STIS
Jakarta, Indonesia
ak.monika@stis.ac.id

Nucke Widowati K. Projo
Data Science Research Unit
Politeknik Statistika STIS
Jakarta, Indonesia
nucke@stis.ac.id

***Abstract*—Indonesia used a Standard Classification of Indonesian Business Fields (KBLI) to classify and categorize economic activities and businesses based on their primary functions and outputs. Practically, to determine the economic activity of a worker, a questionnaire is designed using open-ended questions that generate free-form text responses, and then the enumerator manually selects the appropriate category of KBLI. In this study, we develop text classification models using machine learning and transfer learning approach to automatically assign the category of KBLI from respondents' economic activity descriptions through free-form text responses. The dataset consists of a pair of respondents' job descriptions and the category of KBLI derived from the tourism survey. Based on the evaluation, our transfer learning model from pre-trained IndoBERT trained on a large monolingual Indonesian corpus outperformed machine learning models but with slight differences.**

***Index Terms*—text classification, KBLI, business fields, transfer learning, machine learning**

I. Introduction

Classification systems play a fundamental role in official statistics by providing structure and consistency to data collection, analysis, and reporting. They are essential tools for policymakers, researchers, and the public to understand and make informed decisions based on accurate and comparable statistical information.

Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) or translated as "Standard Classification of Indonesian Business Fields" is a system used in Indonesia to classify and categorize economic activities and businesses based on their primary functions and outputs. It is an essential tool for the Indonesian government, businesses, and statistical agencies to collect, analyze, and report economic data consistently and accurately. It provides a standardized framework for organizing economic activities and helps identify and understand the composition of various industries in the country. The system is regularly updated and revised to reflect changes in the economy and emerging business fields. Each revision brings improvements and adjustments to ensure the classification remains relevant and up-to-date with current economic trends. The classification system consists of a hierarchical structure similar to the International Standard Industrial Classification (ISIC). It has several levels of classification. Sections are the highest level of aggregation, grouping economic activities into 21 broad categories, such as agriculture, manufacturing, construction, wholesale and retail trade, and others.

The second level is called "divisions." Each section is further divided into divisions, representing more detailed groupings of economic activities. The third level is "groups." Divisions are further subdivided into groups, providing even more specific categories of economic activities. Moreover, the last level is called "classes." Each group is divided into classes representing individual economic activities at the most detailed level.

Determining the classification of business fields in an economic activity study that does not have its own industry classification, such as tourism, is an urgent matter. The tourism sector is an example of economic activity, which consists of several economic activities within the KBLI. Therefore, determining the correct KBLI will greatly assist researchers in analyzing economic activities that are not explicitly stated in the KBLI.

Incorporating KBLI codes into a survey to produce official statistics involves careful planning and execution to ensure accurate data collection and classification of economic activities. Designing a questionnaire to implement KBLI involves carefully crafting questions that elicit relevant information about the economic activities of the survey respondents. The questionnaire should be designed to collect data that can be easily mapped to specific KBLI codes to classify the economic activities accurately.

Therefore, to determine the economic activity of a worker, a questionnaire is designed using open-ended questions that generate free form text responses. Then the response is matched with the description from the classification code. The use of open questions to determine the classification of economic activities is very helpful in determining the classification of business fields in a study whose economic activities do not have an industrial classification. For example the tourism industry. The tourism industry is an economic activity consisting of several classifications of economic activities in the KBLI. The use of open questions in the questionnaire is expected to

979-8-3503-8129-0/23/$31.00 ©2023 IEEE

1. Introduction

- Incorporating KBLI codes into a survey to produce official statistics involves careful planning and execution to ensure accurate data collection and classification of economic activities.
- The use of open ended questions to determine the classification of economic activities is very helpful in determining the classification of business fields in a study whose economic activities do not have an industrial classification.
- Regular survey done by BPS such as SAKERNAS, SUSENAS, other economic survey used this type of questionnaire.
- SIBAKU mobile, help the work of supervisors determine this KBLI code. It provides five-digit KBLI code recommendations based on the keywords entered by the user.
- SIBAKU mobile still has drawbacks: the keywords in search process are “exact matching,” which means that if the keywords provided do not precisely match the collections in the database, SIBAKU mobile cannot provide the expected job classification recommendations.

1. Introduction

- Riset III PKL Polstat STIS studied about tourism worker.
- The tourism industry is an economic activity consisting of several classifications of economic activities in the KBLI. The use of open questions in the questionnaire is expected to provide information about workers who work in the tourism industry
- The challenge in processing free-form text responses is to classify respondents' answers into KBLI. Information from respondents could make the KBLI classification of workers imprecise code

1. Introduction

The purpose of this study is to develop text classification models using machine learning and transfer learning approach to automatically assign the category of KBLI from respondents' economic activity descriptions through free-form text responses.

2. Related Works

developed a question and-answer application to determine the KBLI code for business units or companies that have difficulty determining the business field code or main work field.

Compares two text classification models, Cat boost and Double Random Forest (DRF), to carry out classifications with three class distance categories: high, medium, and low.

have implemented random forest, K-nearest neighbour and logistic regression as classification algorithms for the Text Classification

comparing three supervised ML algorithms that used Bag-of-Words features to detect grammatical ambiguity in software requirements: support vector machine (SVM), random forest (RF), and k-nearest neighbours (KNN).

3. Methodology

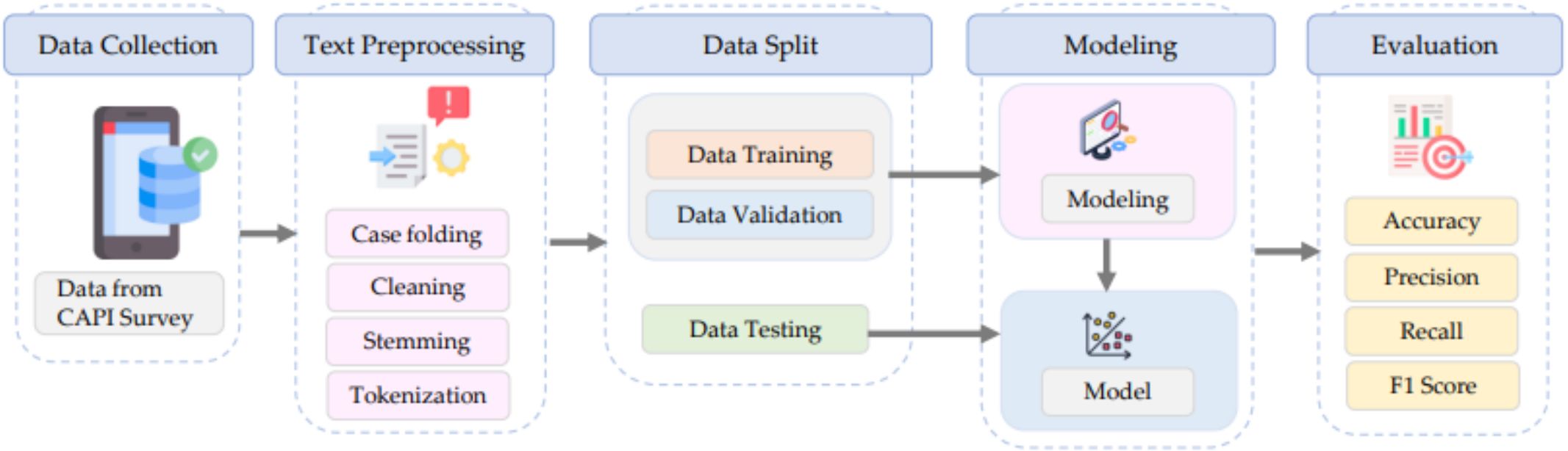


Fig. 1. Research flow.

Methodology – Data Collection

RAHASIA

PRAKTIK KERJA LAPANGAN POLITEKNIK STATISTIKA STIS

LISTING ANGGOTA RUMAH TANGGA YANG TERMASUK

TENAGA KERJA DI INDUSTRI PARIWISATA

VPKLD4-62.LR3

BLOK V. HASIL PENDAFTARAN RUMAH TANGGA (LANJUTAN)

No Urut Rumah Tangga	Keterangan ART Usia 10 Tahun Ke Atas		Apakah Seminggu Yang Lalu Bekerja?	Sementara Tidak Bekerja Karena Cuti, Sakit dan Kondisi Lainnya *[Isikan jika kolom 512 = "2"]*	Apakah Selama Tiga Tahun Terakhir Pernah Bekerja? *[Isikan jika kolom 513 = "2"]*	Lapangan Usaha Saat Ini *[Isikan jika kolom 512 = "1" atau R.513 = "1"]*		Apakah Tempat Bekerja [Nama] Melayani Wisatawan? *[Isikan jika kolom 516 sesuai kode lapangan usaha pariwisata]*	Apakah Selama Tiga Tahun Terakhir Pernah Bekerja Dalam Lapangan Usaha Pariwisata? *[Isikan jika 516 tidak sesuai kode lapangan usaha pariwisata atau kolom 517 = "2"]*	Lapangan Usaha Tiga Tahun yang Lalu *[Isikan jika kolom R.514 = "1" atau 518 = "1"]*		Apakah Tempat Bekerja [Nama] Melayani Wisatawan? *[Isikan jika kolom 520 sesuai kode lapangan usaha pariwisata]*
	No. Urut ART	Nama	1. Ya → **Kolom 515** 2. Tidak	1. Ya → **Kolom 515** 2. Tidak	1. Ya → **Kolom 519** 2. Tidak → **Non eligible, STOP**	Deskripsikan Pekerjaan Responden Saat Ini	Kode Kategori Lapangan Usaha (Diisi Langsung Oleh Petugas) → **Kolom 518 Jika Tidak Sesuai Dengan Lapangan Usaha Pariwisata**	1. Ya → **Eligible, STOP** 2. Tidak	1. Ya → **Kolom 519** 2. Tidak → **Non Eligible, STOP**	Deskripsikan Pekerjaan Responden Tiga Tahun yang Lalu	Kode Kategori Lapangan Usaha (Diisi Langsung Oleh Petugas) → **Non Eligible, STOP Jika Tidak Sesuai Dengan Lapangan Usaha Pariwisata**	1. Ya → **Eligible, STOP** 2. Tidak → **Non Eligible, STOP**
[504]	[510]	[511]	[512]	[513]	[514]	[515]	[516]	[517]	[518]	[519]	[520]	[521]

Methodology – Data Split

Methodology – Text Pre-processing

01 Case folding: converting all letters in the corpus to lowercase.

02 Cleaning: eliminate unnecessary characters such as punctuation and extra spaces.

03 Stemming: removing the inflection of a word into its basic form.

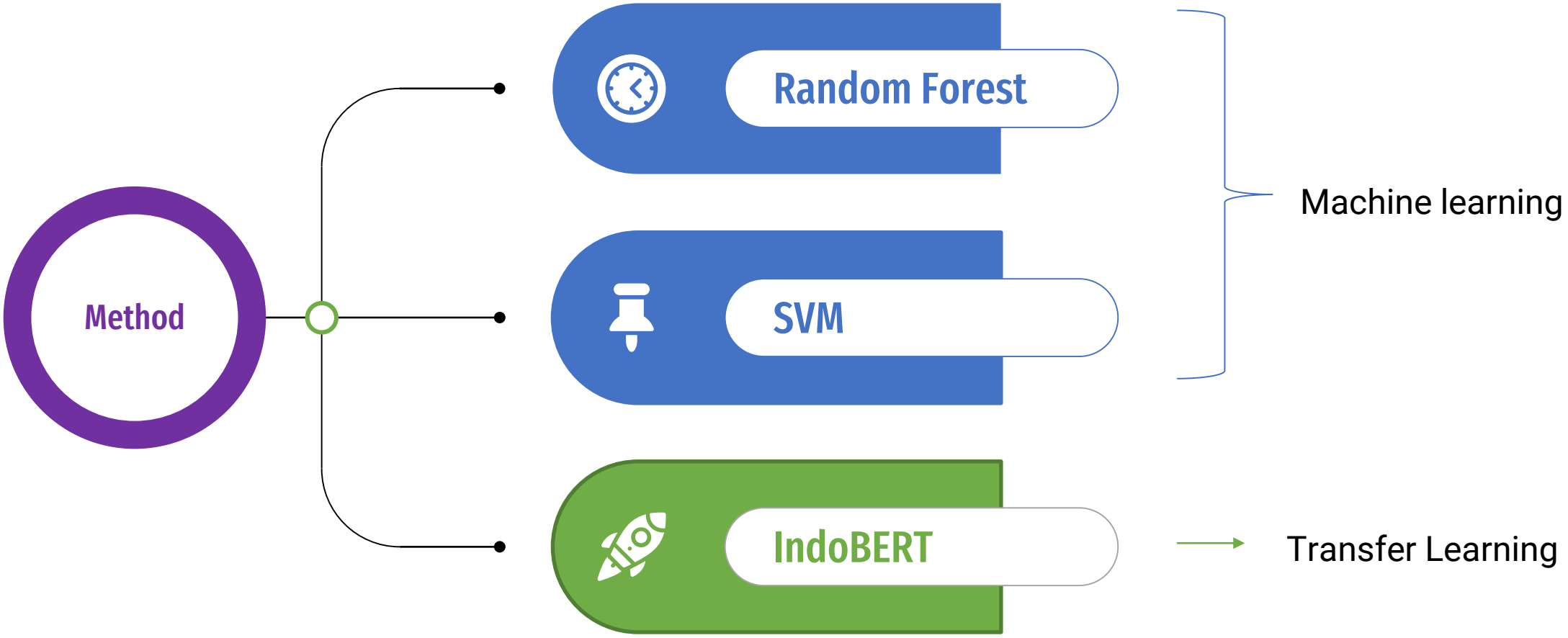
04 Tokenization: separating text into pieces called tokens.

Process	Results
Initial Text	PETANI SAYUR, DI LADANG SENDIRI *VEGETABLE FARMERS, IN OWN FIELDS*
Case Folding	petani sayur, di ladang sendiri *vegetable farmers, in own fields*
Cleaning	petani sayur di ladang sendiri *vegetable farmers in own fields*
Tokenizing	['petani', 'sayur', 'di', 'ladang', 'sendiri'] ['farmers', 'vegetable', 'in', 'fields', 'own']
Stemming	['tani', 'sayur', 'di', 'ladang', 'sendiri'] ['farm', 'vegetable', 'in', 'fields', 'own']

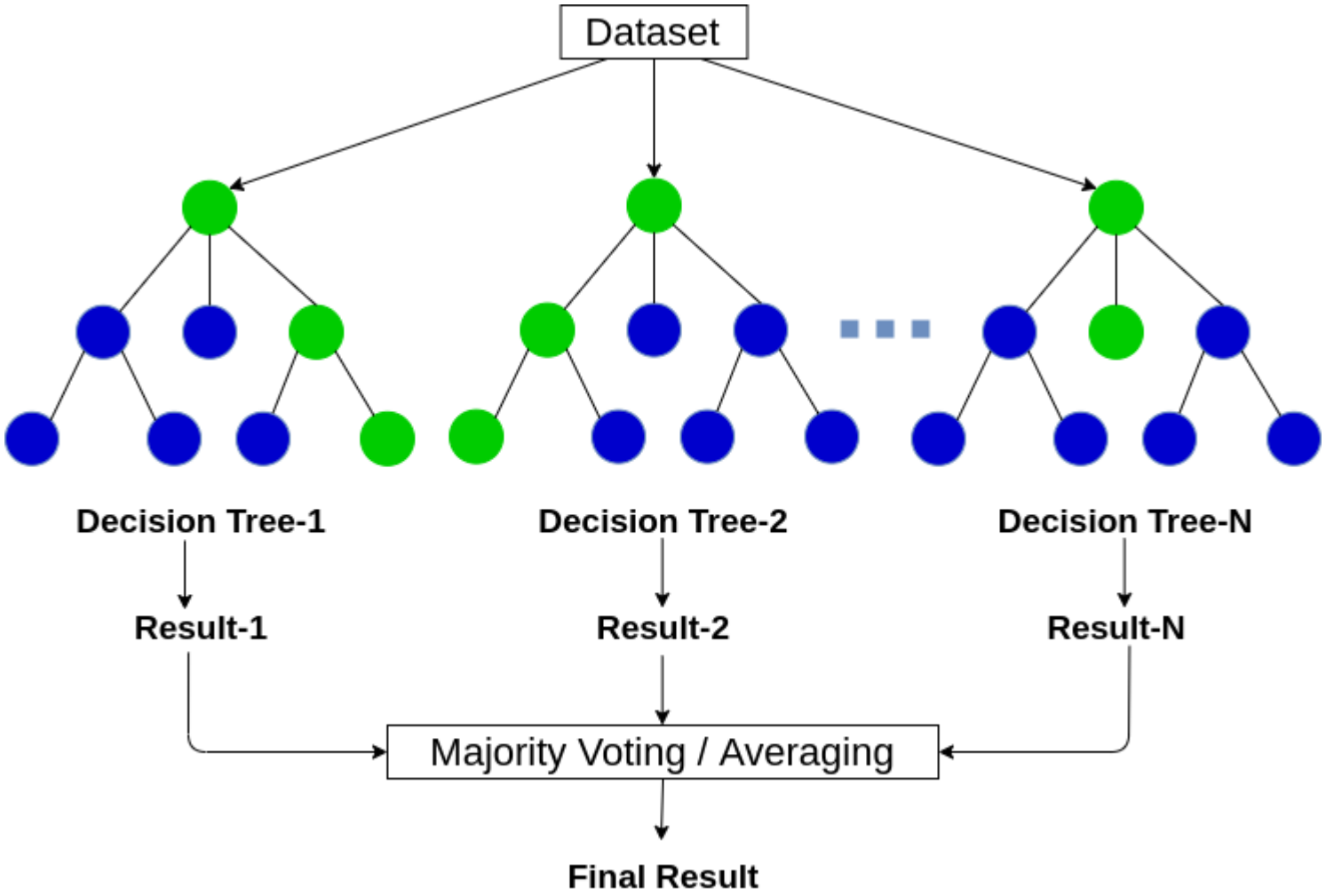
Feature Extraction:

- Term Frequency – Inverse Document Frequency (TF.IDF) -> machine learning
- BERT embedding -> transfer learning

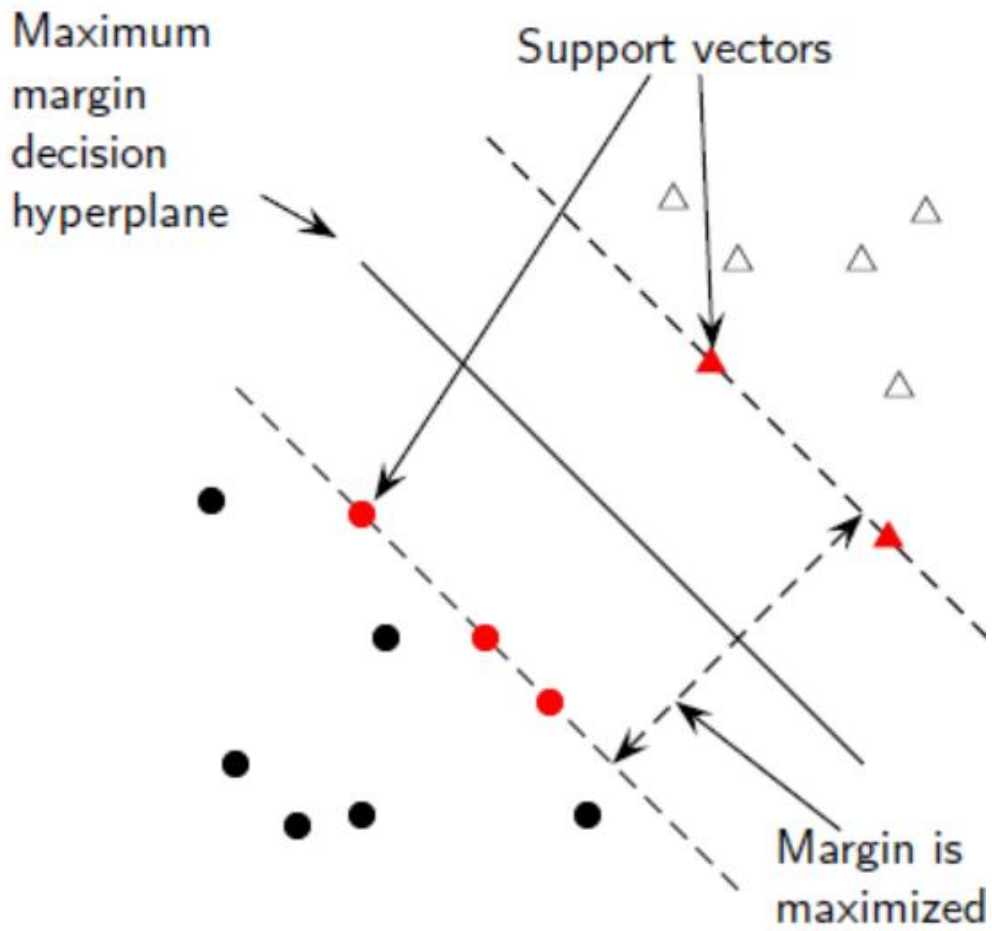
3. Methodology - Text Classification



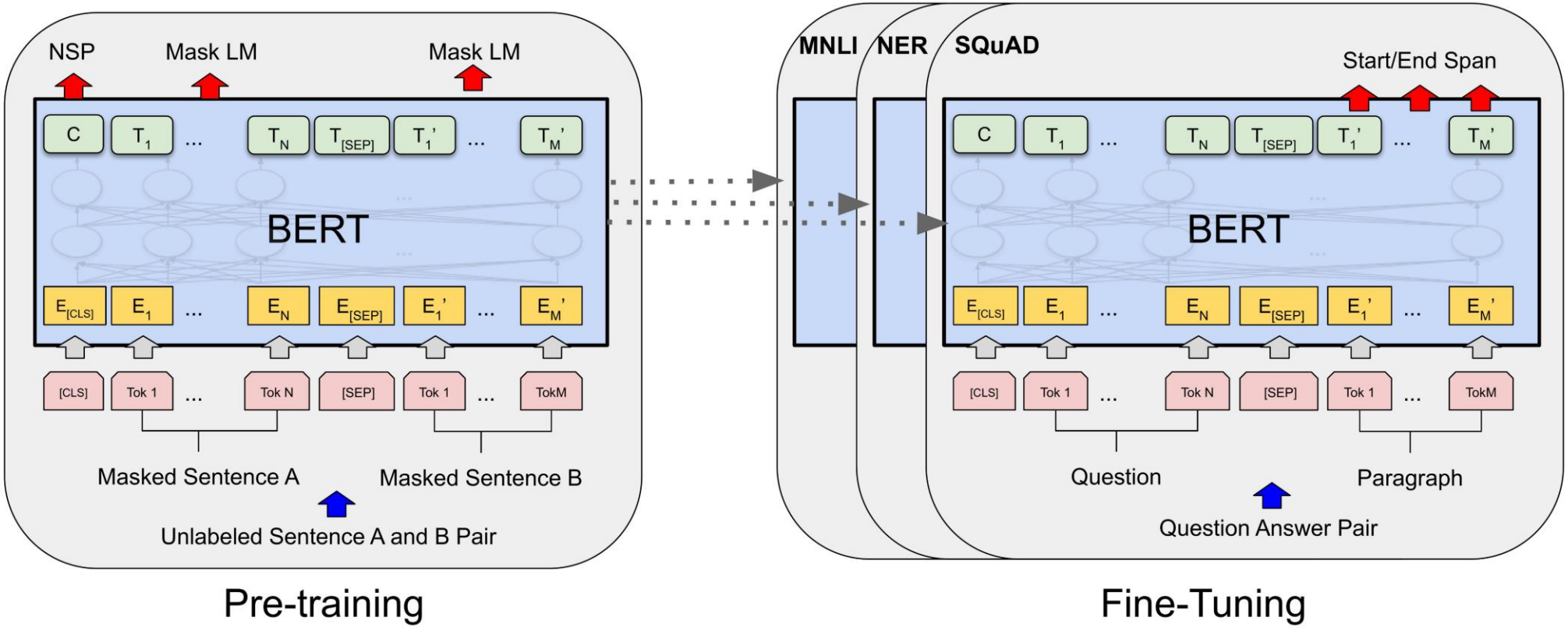
Random Forest



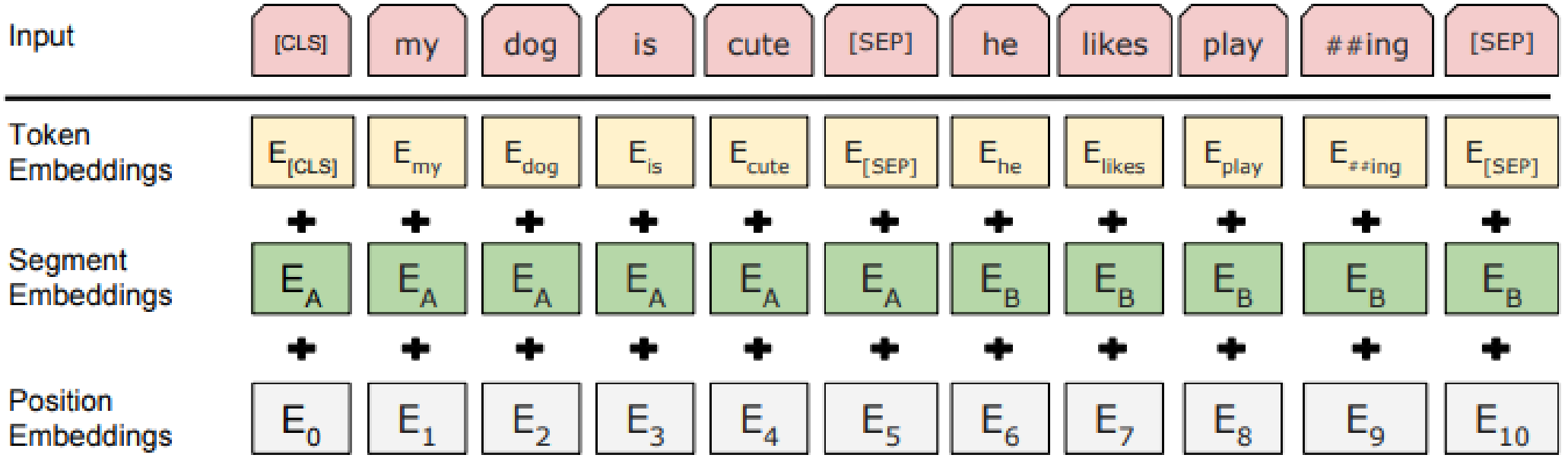
SVM (Support Vector Machine)



BERT
(Bidirectional Encoder Representations from Transformers)



BERT
(Bidirectional Encoder Representations from Transformers)



BERT vs IndoBERT

- BERT trained on BooksCorpus (800M words) and English Wikipedia
- IndoBERT trained from a large and clean Indonesian dataset (Indo4B) collected from publicly available sources such as social media texts, blogs, news, and websites.

Dataset	# Words	# Sentences	Size	Style	Source
OSCAR (Ortiz Suárez et al., 2019)	2,279,761,186	148,698,472	14.9 GB	mixed	OSCAR
CoNLLu Common Crawl (Ginter et al., 2017)	905,920,488	77,715,412	6.1 GB	mixed	LINDAT/CLARIAH-CZ
OpenSubtitles (Lison and Tiedemann, 2016)	105,061,204	25,255,662	664.8 MB	mixed	OPUS OpenSubtitles
Twitter Crawl[2]	115,205,737	11,605,310	597.5 MB	colloquial	Twitter
Wikipedia Dump[1]	76,263,857	4,768,444	528.1 MB	formal	Wikipedia
Wikipedia CoNLLu (Ginter et al., 2017)	62,373,352	4,461,162	423.2 MB	formal	LINDAT/CLARIAH-CZ
Twitter UI[2] (Saputri et al., 2018)	16,637,641	1,423,212	88 MB	colloquial	Twitter
OPUS JW300 (Agić and Vulić, 2019)	8,002,490	586,911	52 MB	formal	OPUS
Tempo[3]	5,899,252	391,591	40.8 MB	formal	ILSP
Kompas[3]	3,671,715	220,555	25.5 MB	formal	ILSP
TED	1,483,786	111,759	9.9 MB	mixed	TED
BPPT	500,032	25,943	3.5 MB	formal	BPPT
Parallel Corpus	510,396	35,174	3.4 MB	formal	PAN Localization
TALPCo (Nomoto et al., 2018)	8,795	1,392	56.1 KB	formal	Tokyo University
Frog Storytelling (Moeljadi, 2012)	1,545	177	10.1 KB	mixed	Tokyo University
TOTAL	3,581,301,476	275,301,176	23.43 GB		

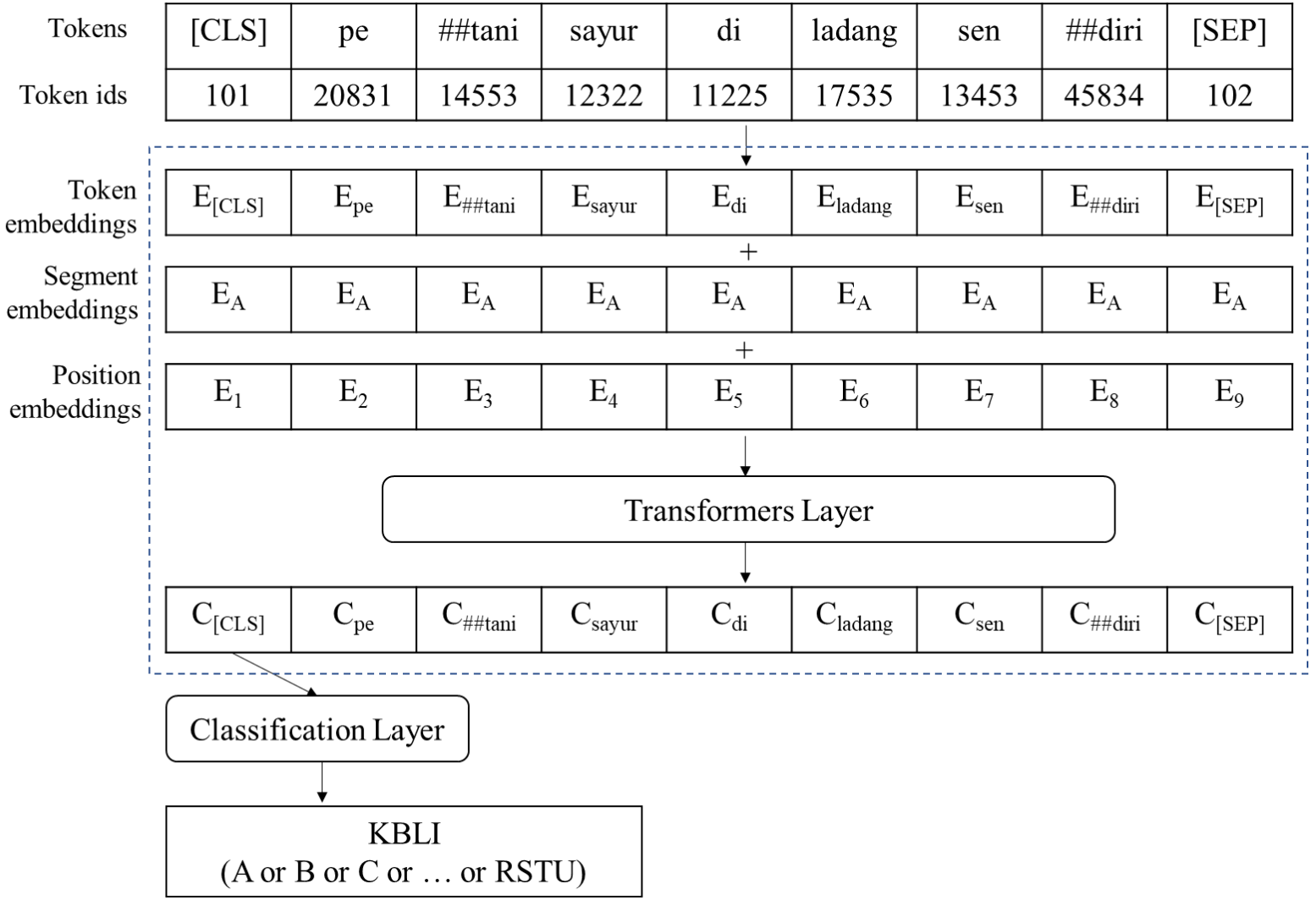
4. Result and Discussion

TABLE II
CATEGORY DISTRIBUTION

KBLI	Total	KBLI	Total	KBLI	Total
A	8,149	G	3,949	MN	272
B	62	H	635	O	293
C	1,248	I	1291	P	765
D	37	J	55	Q	211
E	74	K	95	RSTU	1,398
F	1,388	L	22		

4. Result and Discussion

Fine-tuned IndoBERT Architecture for Businesss Field Categorization



4. Result and Discussion

TABLE III
EXPERIMENTAL RESULTS OF BUSINESS FIELD CLASSIFICATION MODELS

Method	Precision	Recall	F1-Score	Accuracy
Machine Learning				
Random Forest	**0.78**	0.63	0.67	0.86
SVM	0.69	0.63	0.65	0.85
Transfer Learning				
Fine-tuned IndoBERT	0.72	**0.67**	**0.68**	**0.87**

4. Result and Discussion

Error Analysis

The fine-tuned IndoBERT can capture the sentence contexts better than random forest. For example, the fine-tuned model can correctly classify description of “pengiriman sayur ke luar kota” (delivery of vegetables out of town) as Transportation and Storage category (H).

TABLE IV
EXAMPLES OF JOB DESCRIPTIONS WITH PREDICTED AND TRUE LABELS USING RANDOM FOREST AND INDOBERT

No	Job Description	Predicted Label of Random Forest	Predicted Label of In-doBERT	Actual Label
1.	buruh tani di lahan orang lain *farm workers on other people’s land*	A	A	A
2.	buka toko roti di malang untuk ole-oleh *open a bakery in Malang for souvenirs*	A	G	G
3.	pengiriman sayur ke luar kota *delivery of vegeta-bles out of town*	A	H	H

5. Conclusion

- The fine-tuned IndoBERT models trained on a large monolingual Indonesian corpus outperformed machine learning models but with slight differences.
- It indicates that our fine-tuned model still cannot exploit their context-understanding ability from the short text of the job description.
- For further research, it is recommended to develop a text classification model that is robust and effective for short text.

THANK YOU