
JLBert: Japanese Light BERT for Cross-Domain Short Text Classification

LREC COLING 2024

**Chandrai Kayal, Sayantan Chattopadhyay, Aryan Gupta, Satyen Abrol, Archie Gugol
Rakuten Institute of Technology, Japan**

Contents

- Introduction
- Scope of Work
- Phases of Model
 - ◆ Model Focus Area
 - ◆ Model Pre-Training
 - ◆ Methodology
 - ◆ Model Characteristics
 - ◆ Model Fine-Tuning
 - ◆ Fine-Tuning Performance
- Short Text Classification with Large Language Models
- Summary

Pre-Trained Language Models (PLMs)

What are PLMs?

- Trained on a large corpus of text data.
- Understand and generate human language by learning the statistical patterns in the data.
- Leverage the knowledge gained from initial training to perform specific tasks (text classification, sentiment analysis, machine translation etc.)

Challenges:

- Computationally heavy models
- Mostly PLM training data consists of: Wikipedia and Common Crawl corpus
- Model Inferencing requires significant computational resources.
- Majority of PLMs are trained on English language.

Language: English vs Japanese Language



Writing System: English uses an alphabetic writing system, whereas Japanese uses a combination of three scripts: Kanji, Hiragana, and Katakana.



Word Boundaries: English has clear boundaries between words, indicated by spaces. In contrast, Japanese writing does not use spaces, making it harder to identify individual words.



Grammatical Structure: English follows a Subject-Verb-Object (SVO) word order, whereas Japanese typically uses a Subject-Object-Verb (SOV) order.

Hence PLMs used for English language cannot be used for Japanese because of structural differences in both languages.

What are Short Texts?

- Short text refers to textual data that is brief and concise composed of few words or sentences.
- Character length of sentence is not more than 100 characters.
- **Short Text Classification (STC) Applications:** Tweet Categorisation, News Headline Classification etc.
- Despite being short in nature they convey a variety of sentiments, opinions, or information.

Examples of Japanese short texts:

1. Review Titles:

- a. "まあまあ" (So-so)
- b. "残念な結果" (Disappointing result)

2. Tweets:

- a. "今日の天気は最高だね" (Today's weather is great)
- b. "この本、おすすめです！" (I recommend this book!)

3. Headlines:

- a. "地震、被害拡大" (Earthquake, damage expands)
- b. "オリンピック、金メダル獲得" (Olympics, gold medal won)

Scope of Work

- Lack of lightweight models for Japanese texts
- Limited models trained on diverse datasets
- Improving Short Text Classification (STC) tasks performance in Japanese language

To solve this, we propose JLBert

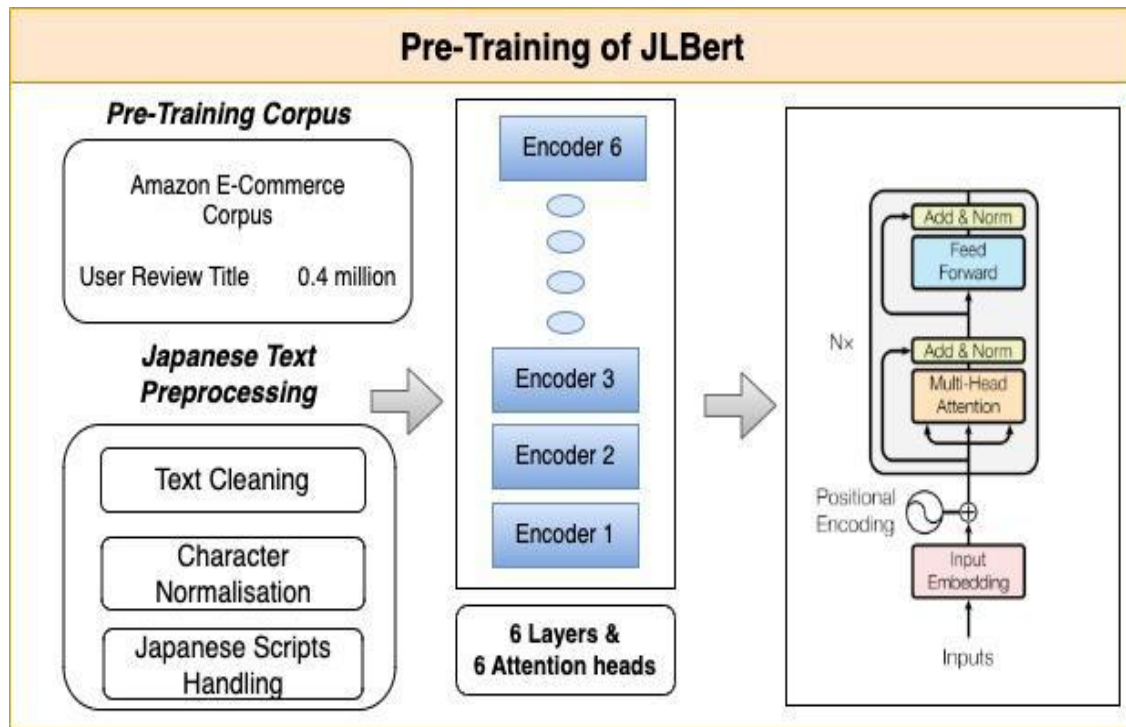
novel cross-domain compact and scalable model, specifically designed for Short Text Classification tasks on Japanese language

JLBert Focus Areas

- 1. Computational Efficiency:** JLBert is designed with computational cost-effectiveness in mind, making it a "greener" choice of modelling. By reducing the number of layers and attention heads (6 layers-6 attention heads), JLBert emerges as a lightweight, fast, and scalable solution, therefore reducing fine-tuning time by 50% across various datasets.
- 2. Proficiency in STC tasks:** JLBert is pre-trained on Amazon User Review Titles to focus on STC tasks. Ecommerce data is a rich and underutilized source of short text data that provides a more consumer-centric perspective and offers various insights into user sentiment, preferences, and language usage.
- 3. Cross-domain applicability:** Despite its pretraining on a specific domain (e-commerce) data, JLBert demonstrates fine-tuning adaptability across various domains such as fintech, news, etc

In summary, the novelty of JLBert lies in its unique pre-training on short texts, its cross-domain nature, and its focus on lightweight "greener" model.

JLBERT Pre-Training:



Features:

1. JLBert is pre-trained on the **Amazon Japanese User Review corpus**.
2. The training dataset consists of **review titles (short texts)**.
3. Training corpus has a combination of hiragana, katakana, kanji & English characters.
4. This is a **Cross Domain model**, trained on e-com data but can be used for other domain downstream tasks.
5. **Greener Light Weight model** as we reduced the size of BERT from 12-layers-12-attention heads to **6-layers-6-attention heads**.

METHODOLOGY

Pre-Training Dataset

- Built a training corpus of **short texts**.
- **0.4M Review Titles** from the Japanese Amazon User Review where the character length of each review title is maximum 20 characters.
- Corpus contains **rich and diverse vocabulary** (books and electronics to clothing and food items)
- Corpus written in 3 scripts: **Hiragana, Katakana and Kanji**.

Pre-Processing

- **Text Cleaning:** Removed Html tags, special characters and japanese stop words.
- **Punctuation Normalization:** Special symbols and numerals (Zenkaku or Hankaku). Used neologdn.
- **Character Normalization:** Convert all characters into a standard form. Used Normalization Form KC (NFKC).
- **Handling Japanese scripts:** Katakana-Hiragana conversion to standardize the japanese texts.

Training

- **Tokenization Strategy** - Byte-Pair Encoding (BPE)
- 6 hidden layers, 6 attention heads, and 768 hidden sizes (lighter than BERT)
- Vocabulary size is 30,522
- Parameters: 13M
- 4 Tesla V100 GPUs
- 4 * V100 * 8hours

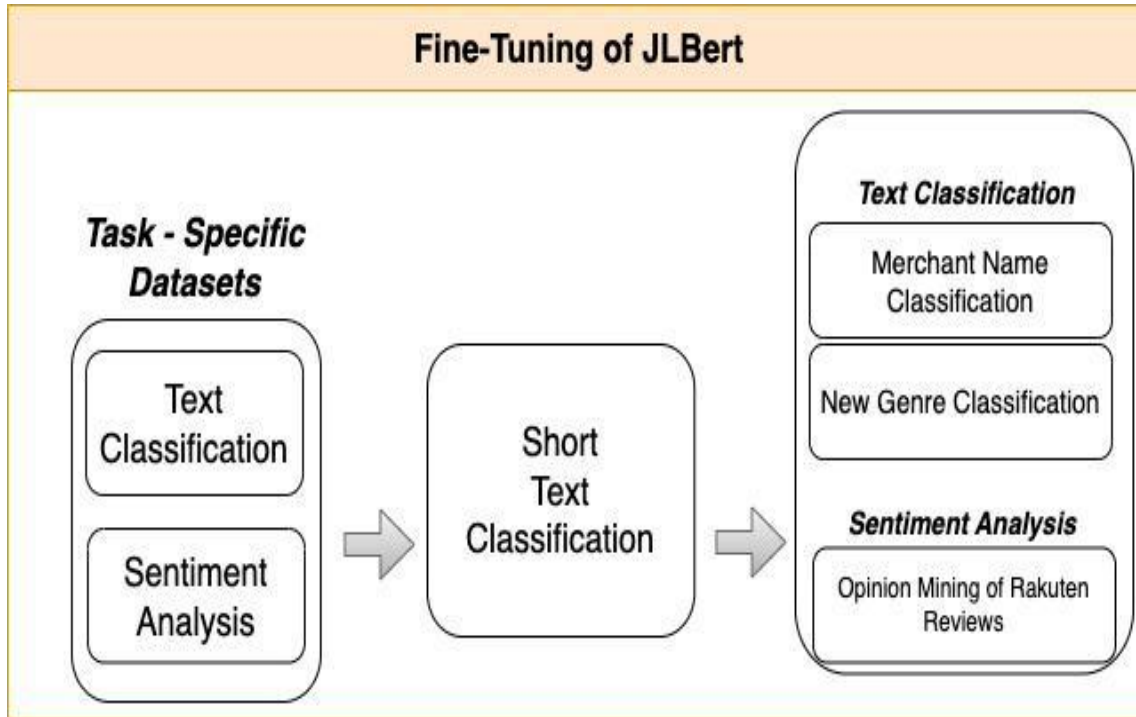
Characteristics of JLBert with other SOTA BERT models

Comparison	mBERT	JapBERT	mDistilBERT	JLBert
Parameters	110M	110M	66M	13M
Layers/ Hidden Dimensions/ Attention Heads	12/768/12	12/768/12	6/768/12	6/768/6
Pre-Training Data	BooksCorpus + Wikipedia	Japanese Wikipedia	BooksCorpus + Wikipedia	Japanese Amazon Reviews
Training Time	4 TPUs*4 days	1 TPU*5 days	8*V100*3.5 days	2*V100*8hours
Method	Transformer, MLM and NSP	BERT without NSP	BERT-Distillation	BERT without NSP

Considered following SOTA models as our baselines.

1. BERT Multilingual model (mBERT)
2. Japanese BERT model (JapBERT)
3. DistilBERT model (mDistilBERT)
4. Japanese Distil- BERT model (JapDistilBERT)

JLBert Fine-Tuning



Fine-tuned on 1 industry + 2 open-source datasets

Transfer-Learning approach for fine-tuning on 3 Short Text datasets:

- ➔ **Merchant Name Classification (MNC)**
 - ◆ Credit Card Merchant categorization model.
 - ◆ Industry dataset for predicting genres from merchant name.
 - ◆ Highly imbalanced data with 23 class classification.
- ➔ **Classification of Japan NHK shows**
 - ◆ News dataset genre classification model from short titles of a show.
 - ◆ Contains 21,795 different show titles with 13 classes.
- ➔ **Sentiment Analysis**
 - ◆ Rakuten review titles is sentiment analysis dataset.
 - ◆ Contains 40k review titles with five user polarity classes.

Fine-Tuning Performance

Datasets	Metrics	JLBert	mBERT	JapBERT	mDistilBERT	JapDistilBERT
MNC	F1-Score	0.8223	0.8154	0.8092	0.8124	0.6250
	CO2 Emission	0.11723	0.39794	0.35720	0.28244	0.32319
	Runtime	45 mins	91 mins	87 mins	56 mins	58 mins
Japan NHK	F1-Score	0.7268	0.7166	0.7014	0.7101	0.5409
	CO2 Emission	0.05076	0.18375	0.17505	0.16745	0.17469
	Runtime	20 mins	45 mins	41 mins	38 mins	40 mins
Rakuten review titles	F1-Score	0.7400	0.7320	0.7380	0.7335	0.7082
	CO2 Emission	0.10351	0.42562	0.36824	0.29826	0.32561
	Runtime	40 mins	98 mins	92 mins	55 mins	60 mins

- JLBert is energy-efficient, resulting in **least CO2 emissions**.
- JLBert outperforms SOTA BERT models by **approx 1.5%**.
- **Runtime for JLBert is minimal** in comparison to others for both fine-tuning and inferencing tasks across all datasets.

Short Text Classification with LLMs

- Performed **Zero-shot, One-shot, and 15-shot classification** strategies with LLMs.
- Models used:
 - **Meta's Llama-2**
 - **Open AI's GPT- 3.5**
 - **Open AI's GPT-4.**
- Models evaluated on the MNC industry dataset.

Prompt Engineering:

- 20+ iterations, utilizing different templates.
- 1st approach -> entire prompt in English, with predefined genres designated for classification.
- 2nd approach -> entire prompt in Japanese.
- 3rd approach -> Combined prompts were composed of a mixture of English and Japanese
- 3rd approach proved to be the most effective prompt.

Methods	Llama-2	GPT-3.5	GPT-4
Zero-Shot	0.362	0.528	0.641
One-Shot	0.371	0.527	0.641
Few-Shot	0.377	0.538	0.698

Performance comparison of LLMs on MNC dataset for 2000 merchant names. **JLBert F1-score for 2000 merchant is 0.855**

Few shot classification with LLM models underperform when compared with JLBert by approx 15%.

Experiment suggests these models may require substantial fine-tuning to effectively handle datasets like MNC

Summary

- Developed compact model specifically designed for the Japanese language.
- JLBert is pre-trained on short-text e-commerce domain data, an area that has been relatively unexplored.
- **Advantages:**
 - JLBert outperforms SOTA BERT models by approx 1.5%.
 - JLBert reduces pre-training time compared to other BERT models.
 - JLBert significantly shortens time required for fine-tuning on various NLP tasks and inferencing.
 - Beneficial for industry-level applications.
- **Development Motivation:** JLBert was developed to meet the research community's need for a smaller, faster Japanese language model.
- **Efficiency:**
 - Usage of compact and efficient model saves time and energy,
 - Allow researchers and industry professionals to quickly utilize this model even without GPU resources.



THANK YOU