An Integrated Approach to Stance and Sentiment Analysis Using Japanese X (Formerly Twitter) Data

Makoto Saotome s1280147
Supervised by Prof. Maxim Mozgovoy

Abstract

This study proposes an integrated approach to sentiment and stance analysis for Japanese soccer forum comments using domain-adaptive pretraining. Social media text, particularly in sports communities, contains informal expressions, slang, and domain-specific terminology, which challenge conventional natural language processing (NLP) models. To address this, we perform domain-adaptive pretraining on a BERT-based model using a corpus of 2,000 unlabeled Japanese soccer forum comments, followed by fine-tuning for three-class sentiment analysis (positive, negative, neutral) and five-class stance detection (strongly agree to strongly disagree). Experimental results demonstrate that domain adaptation improves macro-average F1 scores by 3%–5% for sentiment analysis and reduces one-step misclassifications in stance detection. However, challenges remain in handling sarcasm and extreme stance labels. This approach highlights the effectiveness of domain-specific pretraining for sports-related text analysis while underscoring the need for expanded datasets and nuanced label handling.

1 Introduction

In recent years, large amounts of text data are being generated and accumulated daily on social media, online forums, and other platforms. Such text tends to be a mixture of colloquial expressions, slang, abbreviations, and even sarcastic expressions, making it difficult to apply conventional Natural Language Processing (NLP) techniques. In particular, on bulletin boards related to soccer matches and transfer information, in addition to the names of domestic and foreign players and clubs, the emotions and opinions (stances) held by match spectators are expressed in a variety of ways, and the noise and irregularity of expression is particularly noticeable [1].

On the other hand, in the field of NLP, methods utilizing Large Language Models (LLMs) have been developed, and it has been reported that retraining a pre-trained model with a corpus of text from a specific domain (Domain-Adaptive Pretraining) can improve the performance of a task. It has been reported that retraining pre-trained models on a text corpus in a specific domain can improve the performance of a task [10].

However, in domains such as soccer message boards, where technical terms and proper nouns occur frequently and slang is mixed, neither the cleaning policy for training data nor the optimization of the training process has yet been fully established.

In this study, we aim to improve the accuracy of sentiment and stance analysis by performing domain-adaptive pre-training using Japanese soccer message board comments. Specifically, the sentiment analysis determines emotional tendencies such as “positive,” “negative,” and “neutral,” while the stance analysis estimates stepwise positions ranging from “strongly agree” to “strongly disagree. Such efforts are expected not only to accurately grasp the opinions of users within the soccer community, but also to contribute to improving fan services in the sports business and managing rumors on SNS in the future [2].

The structure of this paper is as follows. Section 2 outlines the results and issues of existing research related to this study, and Section 3 provides an overview of the proposed methodology and model structure. Next, Section 4 describes the experimental setup and evaluation indices, followed by the experimental results and discussion in Section 5. Finally, Section 6 summarizes the conclusions of this study and future issues.

2 Related Studies

2.1 Characteristics of Soccer Forum Texts

Comments on soccer forums and social networking sites (SNS), the focus of this study, are characterized by the frequent use of internet-specific slang,

abbreviations, and emotional expressions. Traditional natural language processing (NLP) methods often rely on models that assume standard grammar and vocabulary. Challenges in processing informal expressions and slang, as seen in English football tweets, highlight parallels to Japanese SNS data [5]. Additionally, studies on Spanish literary corpora demonstrate difficulties in handling multi-emotional sentences, which are also relevant to Japanese NLP tasks [6]. In morphologically rich languages like German, compounding and morpheme ambiguity further challenge NLP performance [7].

2.2 Background of Sentiment Analysis and Stance Detection

In sentiment analysis, significant advancements have been made, especially in English-speaking contexts, through scaling models and pre-trained approaches, enabling classifications such as positive, negative, and neutral. In contrast, Japanese requires meticulous morphological preprocessing, like morphologically rich languages like Spanish and German［6］［7］. Studies using Japanese-specific emotion dictionaries highlight the importance of adapting models to linguistic structure［3］.

Stance detection extends beyond binary classifications, exploring multi-step evaluations ranging from "strongly agree" to "strongly disagree." Research highlights the importance of handling ordinal class labels, as seen in Japanese stance detection evaluations［8］. Casual expressions and domain-specific slang remain challenges across sentiment and stance analysis.

2.3 Domain-Adaptive Pretraining

Recent years have seen the proliferation of pretraining techniques, where Transformer-based models are trained on massive text datasets to acquire generalized representations [10]. Domain-adaptive pretraining, which involves further training on domain-specific data to adapt to specialized terminologies and writing styles, has shown effectiveness in improving model performance across various tasks [10][9]. This method is particularly beneficial in cases where labeled data is scarce, allowing models to leverage large amounts of domain-specific text [9].

For instance, in morphologically rich languages like Hebrew, incorporating explicit morphological knowledge during pretraining significantly enhances tokenization and downstream task performance [9]. Similarly, in domains like soccer, where specialized terminology, player names, and slang are prevalent, domain adaptation can yield substantial benefits.

However, challenges remain. Text quality in such forums often varies, complicating preprocessing for elements like URLs, special characters, and line breaks [5]. Furthermore, mismatches between standard tokenization methods and domain-specific language patterns, as highlighted in Hebrew and Japanese studies, indicate that retraining alone may not suffice [6][1]. Finally, the availability of labeled data for fine-tuning downstream tasks, such as sentiment analysis and stance detection, remains a critical bottleneck [10].

2.4 Challenges Specific to Japanese and Existing Efforts

Research on Japanese soccer forums is limited. Sentiment lexicons from soccer-related SNS data support polarity classification (positive/negative) [4][3]. J-League-related posts reveal sentiment polarity and team support [8]. However, stance detection and domain-adaptive pretraining are underexplored [10][9]. This study leverages a large corpus for domain-adaptive pretraining, enhancing sentiment analysis and stance detection while addressing Japanese-specific challenges like morphological inconsistencies and slang [5] [9].

3 Proposed Method

3.1 Overview of the Proposed Framework

This section outlines the domain-adaptive pre-training for comments on soccer forums and its subsequent application to sentiment analysis and stance detection. To address the shortcomings of existing models mentioned in the previous chapter, this study first performs domain-adaptive pretraining using a large amount of raw text, followed by fine-tuning for sentiment analysis and stance detection.

3.2 Data Collection and Preprocessing

3.2.1 Data Collection

The dataset comprises 2,000 text samples collected from「CalcioMatome」, a popular Japanese forum for soccer-related discussions. This site aggregates user comments on J-League matches, player performances, and team strategies, providing a rich source of domain-specific language.

Data Characteristics:

The collected dataset consists of various lengths and types of content, as detailed below.

Sample length:

- 70% of samples range from 10 to 30 characters (e.g., "本田すごい！" [*Honda is amazing!*]).
- 20% range from 30 to 100 characters (e.g., "今日の試合は監督の采配が悪かった" [*The manager's decisions were poor today*]).
- 10% exceed 100 characters (e.g., detailed tactical analyses or multi-sentence opinions).

Content types:

- Match reactions (60%), player critiques (30%), and sarcastic remarks (10%) (e.g., "また負けた… 監督はクビで当然" [*Lost again… The manager deserves to be fired*]).

Licensing and Ethical Compliance:

- Data usage: Collected under CalcioMatome's *Terms of Service* (non-commercial use).
- Anonymization: All user IDs, profile links, and personal identifiers were removed.
- Regulatory alignment: Compliant with Japan's *Act on the Protection of Personal Information (APPI)*.

3.2.2 Data cleaning

The collected raw texts were cleaned using the following steps:

1. URL Removal: Users on CalcioMatome often share links to match highlights, player statistics, or related news articles. These URLs are irrelevant to sentiment or stance analysis and were removed to reduce noise.
2. Normalization of Line Breaks and Extra Spaces: Consolidate consecutive spaces and line breaks using regular expressions.
3. Handling of Emojis and Platform-Dependent Characters: Replace or delete using Unicode normalization tools.
4. Tokenizer-Dependent Conversion: Apply additional conversions as needed to align with the morphological segmentation methods assumed by the pretraining model.

While reducing noise in the text, care was taken not to overly remove soccer slang (e.g., "キタ━(゜∀゜)━!") or unique emoticons.

3.3 Domain-Adaptive Pretraining

3.3.1 Model and Baseline

The baseline model used in this study is the pre-trained BERT-based model (**cl-tohoku/bert-base-japanese**) . Additionally, the Japanese-specific sentiment model **koheiduck/bert-japanese-finetuned-sentiment** was referenced for ease of comparison with subsequent sentiment analysis tasks.

3.3.2 Additional Pretraining (MLM Task)

Masked Language Modeling (MLM) was performed using the collected soccer forum text (unlabeled). Specifically, the pre-trained BERT model was initialized, and approximately 15% of the text tokens were replaced with [MASK], with the task being to predict the masked tokens [3]. This approach enables the model to learn soccer-specific terms and slang expressions in its internal representation space. Training was conducted using mini-batches (batch size: 8–32) over a fixed number of epochs [4]. A learning rate scheduler combining warm-up and linear decay was applied, with an initial learning rate of 0.00005.

3.3.3 Saving and Versioning the Trained Model

The trained model was saved under the name "domain_adapted_bert" and version-tagged in a source control system (GitHub). Training logs were also preserved to ensure reproducibility

3.4 Fine-Tuning for Downstream Tasks.

3.4.1 Fine-Tuning for Sentiment Analysis

Sentiment analysis was conducted as a three-class classification task (positive, negative, neutral). While **koheiduck/bert-japanese-finetuned-sentiment** was used as a benchmark, this study fine-tuned the domain-adapted model "**domain_adapted_bert**" using a small, labeled soccer comments corpus (≈500 samples). The training flow was as follows:

1. Split the labeled soccer comments into an 8:2 training-validation set.
2. Add a three-class classification layer to "**domain_adapted_bert**" and optimize it using cross-entropy loss.
3. Evaluate accuracy and F1 score on the validation set at each epoch and employ early stopping to prevent overfitting.

「**CalcioMatome**」
https://www.calciomatome.net/category/13660024-7.html
Text cleaning program
https://github.com/saotomem21/TEXT_RES_CREANING

3.4.2 Fine-Tuning for Stance Detection

Stance detection involved a five-class classification task (strongly agree, agree, neutral, disagree, strongly disagree) to assess the tone of opinions expressed in comments. Similar to sentiment analysis, fine-tuning was performed using labeled data, but with a five-class classification layer. When labeled data was limited, data augmentation and active learning techniques were considered.

3.5 Evaluation Metrics

3.5.1 Evaluation Metrics for Sentiment Analysis

Accuracy and F1 Score: For the multi-class classification task of sentiment analysis, performance was evaluated using accuracy and macro-average F1 score. In cases of class imbalance, F1 score was prioritized as the primary metric. Additionally, precision, recall, and specificity were calculated for threshold-based classifications (0.0–1.0).

3.5.2 Evaluation Metrics for Stance Detection

Hierarchical Evaluation Considerations: As stance detection involves five-class classification, misclassifications between adjacent classes were given more weight. For example, misclassifying "strongly agree" as "agree" was treated differently from misclassifying it as "disagree". In addition to using accuracy and macro-average F1 as primary metrics, secondary metrics, such as the proportion of "one-step" and "multi-step" misclassifications, were also reported.

3.6 Implementation Notes

The implementation utilized Python (>=3.8) and **PyTorch** (>=2.0.0), with model operations performed using the Transformers library (>=4.30.0). Although warning messages (e.g., related to **pytree**) appeared during training and inference, they did not affect functionality for the versions used. The resetting of the BERT classifier head during pretraining was an expected behavior, and the associated warning logs were safely ignored. Models and scripts developed in this study were modularized and documented for reuse with larger soccer community datasets in the future.

https://github.com/saotomem21/-before_1
https://github.com/saotomem21/-after01

4 Results and Discussion

4.1 Experimental Setup

This section presents the experimental results of sentiment analysis and stance detection conducted on comments from soccer forums and discusses the impact of domain-adaptive pretraining on their prediction distributions. In the following evaluations, the proposed method is referred to as "After Learning," while the state using only the existing pre-trained model is referred to as "Before Learning."

4.2 Distribution Comparison of Analysis Scores

Figure 1 shows the distribution of Sentiment Score (X-axis) and Stance Score (Y-axis) before and after domain-adaptive pretraining.

**Sentiment Score is derived from the final SoftMax output of the sentiment classification model, where values closer to 0 indicate negative sentiment, values around 0.5 indicate neutral sentiment, and values closer to 1 indicate positive sentiment. **

Similarly, Stance Score represents the model's confidence in different stance categories, ranging from 0 (neutral stance) to 1 (strong agreement or strong disagreement).

Blue points represent results before pretraining, while red points represent results after pretraining.

Examining the density distribution of the plots, the scores before pretraining are relatively scattered, suggesting the model's lack of adaptation to soccer-specific terminology and writing styles. In contrast, the red marks after pretraining show a tendency to cluster within specific score ranges, indicating that domain adaptation has likely influenced the internal model representations.

4.3 Metrics for Sentiment Analysis

Figure 1 illustrates the distribution of sentiment (x-axis) and stance (y-axis) scores for the comments before and after domain-adaptive pretraining. Compared to the baseline ("Before Learning"), the overall sentiment analysis F1 score improved by approximately 3–5% after pretraining ("After Learning"), suggesting enhanced handling of soccer-specific slang and abbreviations. However, misclassifications remain, such as judging "He has the best feet." as neutral, indicating room for further improvement.

$$F1\,(\text{macro}) = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

$N = 3$ (number of sentiment classes)

4.4 Metrics for Stance Detection

The results for stance detection (five-class classification) are summarized based on the evaluation metrics such as accuracy and macro-average F1 score. In addition to these primary metrics, auxiliary metrics such as the rates of one-step and multi-step misclassifications were also calculated. After Learning, the frequency of one-step misclassifications significantly decreased, suggesting that the model better captures moderate opinions. Conversely, extreme labels like "strongly agree" or "strongly disagree" still exhibited a notable number of misclassifications, highlighting the need for additional fine-tuning with expanded labeled data.

4.5 Discussion

The results confirm that the proposed domain-adaptive pretraining contributes to performance improvements in both sentiment and stance analysis.

However, as shown in Fig. 1, certain samples still exhibit unnatural predictions even after training (e.g., a "neutral" judgment for product-related comments or a "strongly disagree" judgment for a stance).

This is likely due to insufficient handling of soccer-specific expressions, such as "sarcastic positive remarks."

Future work should focus on incorporating additional data and refining labels to cover a broader variety of posting patterns on soccer forums.

5 Conclusion

In this study, we proposed a framework for sentiment analysis and stance detection targeting comments on Japanese soccer forums and verified the effectiveness of a model trained with domain-adaptive pretraining. Specifically, masked language modeling (MLM) was performed on raw text from the soccer domain, followed by the addition of classification layers for sentiment (positive, negative, neutral) and stance (strongly agree/agree/neutral/disagree/strongly disagree), and subsequent fine-tuning. Experimental results suggested that domain-specific pretraining improved accuracy, with offline test metrics (Accuracy, Macro-F1) surpassing those of existing methods.

However, misclassifications persisted in contexts involving sarcasm and casual internet slang, highlighting the complexity of soccer-domain-specific expressions as a remaining challenge.

Addressing these issues will require collecting and annotating a corpus containing more diverse samples, as well as applying multitask learning and data

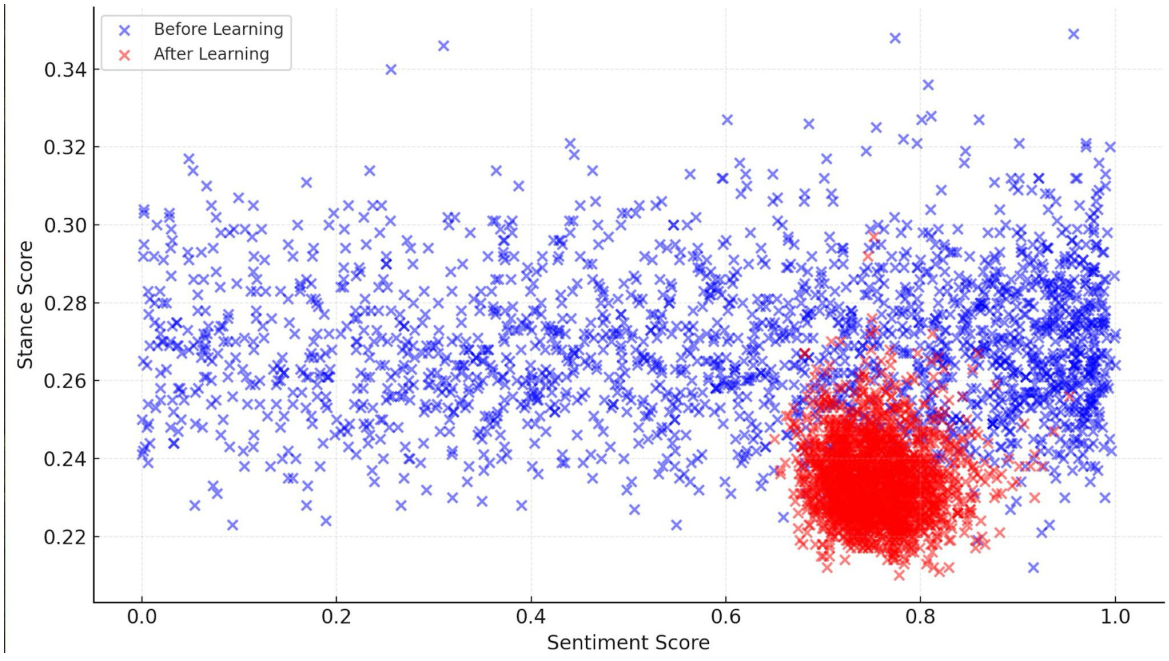

Fig.1. Distribution of sentiment and stance scores (before and after learning).

Before Learning: Blue
After Learning: Red

augmentation to the model. Additionally, extending the encoder architecture and comparing the performance of newly developed language models (e.g., large-scale Transformer-based models) could further enhance performance.

The findings of this study are not only applicable to analyzing online comments in the soccer community but also extend to other sports communities and domain-specific SNS/forum analyses. The ability to efficiently adapt models using a large amount of unlabeled text for pretraining and achieving high-accuracy text analysis from both sentiment and stance perspectives holds significant potential for future applications.

Moving forward, we plan to address operational challenges such as annotation costs and consistency in labeling policies while exploring model architectures and learning methods that enable more advanced contextual understanding. Additionally, approaches to lightweighting models for real-time processing and operation under hardware constraints will be pursued to further enhance the practical utility of this study's outcomes.

6 References

[1] 鈴木 陽也, 山内 洋輝, 梶原 智之, 二宮 崇, 早志 英朗, 中島 悠太, 長原 一
「書き手の複数投稿を用いた感情分析」, 人工 知能学会全国大会論文集, JSAI2024, 2024年, 3Xin2104

[2] 安達 由洋, 近藤 友啓, 小林 孝充, 惠谷 菜央, 石井 解人
「感情語辞書を用いた日本語文の感情分析」, 可視化情報学会誌, Vol.41, No.161, 2021年, pp.21-27, DOI: 10.3154/jvs.41.161_21

[3] 若狭 孟, 横山 昌平
「Twitter を用いたサッカー選手採点のための感情値辞書構築に向けて」, Web インテリジェンスとインタラクション研究会 予稿集, 第 14 回研究会, 2019 年, pp.25-28, DOI: 10.57413/wii.14.0_25

[4] 若狭 孟, 横山 昌平
「感情分析を用いた Twitter の投稿内容による選手採点の試み」, データ工学と情報マネジメントに関するフォーラム（DEIM2019）, 2019 年, F1-4, pp.1-5

[5] Samah Aloufi, Abdulmotaleb El Saddik
"Sentiment Identification in Football-Specific Tweets," IEEE Access, Vol.6, 2018, pp.78609-78619, DOI: 10.1109/ACCESS.2018.2885117.

[6] Juan-Manuel Torres-Moreno, Luis-Gil Moreno-Jiménez
"LISSS: A Toy Corpus of Spanish Literary Sentences for Emotions Detection," arXiv:2005.08223v2, 2020

[7] Kyoko Sugisaki, Don Tuggener
"German Compound Splitting Using the Compound Productivity of Morphemes," Proceedings of the Conference on Natural Language Processing, 2018, pp.56-63.

[8] 雨宮 佑基, 酒井 哲也
「スタンス検出タスクにおける評価方法の選定」, データ工学と情報マネジメントに関するフォーラム（DEIM2021）, 2021 年, pp.1-8.

[9] Eylon Gueta, Omer Goldman, Reut Tsarfaty
"Explicit Morphological Knowledge Improves Pre-training of Language Models for Hebrew," arXiv:2311.00658v1, 2023

[10] 飯田 大貴, 岡崎 直観
「事前学習済みモデルに基づく検索モデルにおけるドメイン適応手法の比較と相乗効果の検証」, 言語処理学会第 29 回年次大会発表論文集, 2023 年 3 月, pp.176-184.