Comparative Analysis of Cross-Lingual Sentiment Classification Models: Evaluating Performance Using the MAD-TSC Dataset

Yiqin Zhou and **Xi Zheng**
University of California, Berkeley
yiqzho@berkeley.edu
azheng0621@berkeley.edu

Abstract

Amidst an increasingly interconnected global milieu, comprehending the intricacies of multilingual dynamics in natural language processing (NLP) emerges as a crucial endeavor. However, prevailing research predominantly fixates on monolingual datasets, thereby neglecting the nuanced intricacies inherent in multilingual communication, particularly within the realm of news discourse. To bridge this lacuna, our study embarks on an exploration of sentiment analysis spanning eight languages, facilitated by the "Multilingual Aligned News Dataset for Target-dependent Sentiment Classification" (MAD-TSC). Harnessing this extensive dataset, we undertake a comprehensive comparative assessment of sentiment classification models, encompassing both the BERT base model and the BERT multilingual base model, across diverse linguistic contexts. Our empirical findings corroborate the efficacy of target-dependent models while also revealing the general effectiveness of MBert in multilingual settings, notwithstanding sporadic performance disparities across different languages. Notably, our analysis underscores the comparatively inferior performance of English languages relative to their counterparts, signifying intriguing linguistic nuances warranting further investigation.

1 Introduction

In the contemporary interconnected global landscape, the significance of conducting multilingual studies cannot be overstated. As our world becomes increasingly interconnected through globalization, immigration, and digital communication, understanding the dynamics of multilingual contexts is imperative. However, despite this importance, much of the existing research in natural language processing (NLP) has predominantly focused on monolingual datasets. This narrow focus overlooks the rich complexities inherent in multilingual communication, including linguistic diversity, cultural nuances, and language-specific phenomena.

In light of this gap in the literature, our research endeavors to address the need for comprehensive multilingual studies. To this end, we have chosen to undertake a comparative analysis using the "Multilingual Aligned News Dataset for Target-dependent Sentiment Classification" (MAD-TSC). This dataset stands out as the first large-scale multilingual aligned dataset tailored specifically for target-dependent sentiment classification in news articles. It offers a unique opportunity to explore sentiment classification across multiple languages, providing professionally-translated and aligned versions of sentences in eight languages, including English, Spanish, German, Italian, French, Portuguese, Dutch, and Romanian. [3]

Our choice of the MAD-TSC dataset is motivated by its comprehensive coverage of languages and its alignment of translated sentences, facilitating consistent analysis across language boundaries. By leveraging this dataset, we aim to investigate how different versions of the BERT model, including the "BERT base model (cased)" trained primarily on English and the "BERT multilingual base model (cased)" trained on a diverse set of languages, perform across various linguistic contexts. [2] Specifically, we are interested in exploring whether the multilingual BERT model outperforms its English-centric counterpart when applied to languages other than English.

Furthermore, our research seeks to delve into the nuanced performance variations of different languages using the same model and methodology. By analyzing the effectiveness of sentiment classification across diverse linguistic datasets, we aim to uncover insights into the challenges and opportunities presented by multilingual NLP tasks. This study holds broader implications for the field of NLP, as sentiment classification serves as a fundamental task that underpins various downstream

applications, including text classification, recommendation systems, chatbots, and more. Through our investigation, we aim to contribute to a deeper understanding of multilingual sentiment classification and its implications for cross-lingual communication and information processing.

2 Background

Previous research has delved extensively into multilingual target-dependent sentiment classification, predominantly within the realms of social media platforms. Studies such as XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond (Barbieri et al., LREC 2022) have contributed significantly to understanding sentiment dynamics across different languages in the context of Twitter [1]. However, our study endeavors to carve a unique niche by shifting the focus towards sentiment analysis in news articles, a domain that remains relatively underexplored yet holds paramount importance as a primary source of information for individuals worldwide. Unlike social media, news articles often exhibit more formal and structured language, posing distinct challenges and opportunities for sentiment analysis.

While the paper introducing the MAD-TSC dataset [3] laid the groundwork for evaluating state-of-the-art Target-dependent Sentiment Classification (TSC) methods in both monolingual and multilingual settings, its primary emphasis was on TSC methodologies. Unfortunately, it did not delve into comparative analyses between "BERT base model (cased)" and "BERT multilingual base model (cased)" (M-Bert). This represents a critical gap in the literature, especially considering the growing interest in leveraging multilingual models for sentiment analysis tasks. Existing studies have highlighted M-Bert's remarkable ability for cross-lingual generalization [4]. However, it's worth noting that performance variations across different languages, particularly between low-resource and high-resource languages, have been observed [5].

Despite the wealth of research on multilingual models, most studies have focused on general language tasks rather than sentiment analysis. Furthermore, many studies have relied on datasets sourced from platforms like Wikipedia, which may not fully capture the nuances of sentiment in real-world contexts. Therefore, our study seeks to address these gaps by exploring the performance of the MAD-TSC dataset using non-TSC models and conducting a comprehensive comparative analysis of the performance between base BERT and M-BERT models. By doing so, we aim to provide valuable insights into the effectiveness of multilingual models for sentiment analysis tasks, particularly in the domain of news articles.

3 Methods

3.1 Dataset

The MAD-TSC dataset, encompassing eight languages in its entirety, serves as the foundation of our study. Each language within the dataset is meticulously curated, comprising labeled training, validation, and test data sets, as outlined in Table 1. For further elucidation regarding this dataset, please refer to Appendix A.

	Negative	Neutral	Positive
Training	1363	1493	954
Validation	112	117	71
Test	364	401	235
Total	**1839**	**2011**	**1260**

Table 1: Descriptive Statistics of MAD-TSC Dataset

3.2 Model

This section provides a concise overview of the methodologies employed for the baseline model, fine-tuned BERT model, target-dependent fine-tuned BERT model, and target-dependent fine-tuned multilingual BERT model.

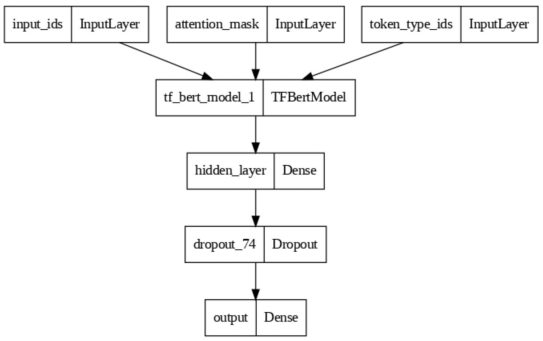
3.2.1 Baseline Model

For this sentiment classification task, we opted for a basic multinomial Naive-Bayes model as the baseline. Multinomial Naive Bayes (MNB), known for its simplicity and efficiency, is widely employed as a standard in NLP, especially for text classification, due to its capacity to represent word occurrences and its interpretability, serving as a useful reference point for evaluating more intricate models.

3.2.2 Fine-tuned BERT Model

Our initial strategy involved deploying a fine-tuned BERT model, leveraging the 'bert-base-cased' architecture. Following the loading of the pre-trained BERT model, we incorporated a hidden layer with Rectified Linear Unit (ReLU) activation and dropout regularization to introduce non-linearity

and prevent overfitting. Subsequently, a dense output layer with softmax activation was introduced to facilitate multi-class classification. The model was then compiled with the Adam optimizer, sparse categorical cross-entropy loss function, and accuracy metric, ensuring its preparedness for training. For a detailed illustration of the model's architecture, please see the diagram below.



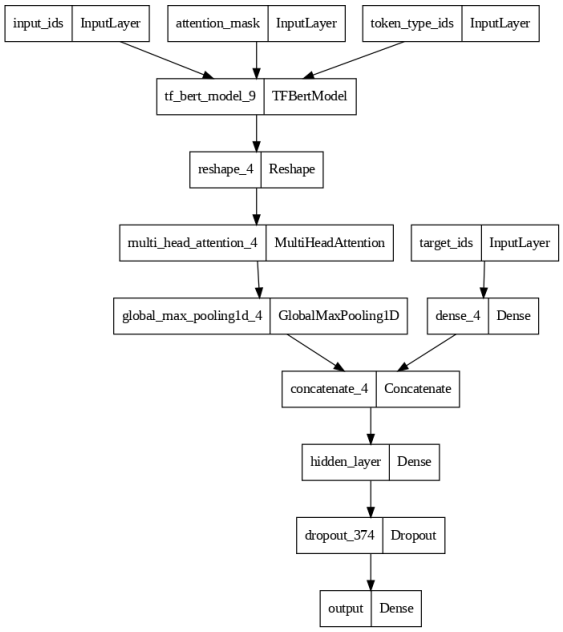
Our approach focused solely on the polarity of the target—negative, neutral, or positive—disregarding any additional target attributes. To streamline training, we imposed a maximum input length constraint of 50 tokens, aiming to expedite training duration while preserving model performance integrity. Through systematic experimentation, we optimized the model's hyperparameters, exploring variations in hidden size, learning rate, batch size, and epochs.

After exhaustive testing, we settled on a configuration comprising a hidden size of 201, dropout rate of 0.3, learning rate of 0.00005, batch size of 8, and 3 training epochs. Due to time constraints, experimentation was restricted to epochs ranging from 1 to 10. Following meticulous evaluation, we determined that employing 3 epochs achieved the most favorable results, striking an optimal balance between training efficiency and model performance.

Regrettably, the model's performance fell short of the baseline standard.

3.2.3 Target-Dependent Fine-tuned BERT Model

The next model we employed is incorporating target information. We used the Our subsequent model integration involved the incorporation of target information, utilizing the same pre-trained BERT model 'bert-base-cased'. Upon obtaining the BERT output, we reshaped it to facilitate the application of a multi-head attention mechanism. Subsequently, global max pooling was applied to extract pertinent features. Following this, a dense layer was introduced to process target IDs, derived from the inputs' target field. The resultant pooled attention output and processed target IDs were then concatenated. The subsequent steps of the process mirrored those of the preceding model. For a detailed illustration of the model's architecture, please refer to the diagram below.



In our model configuration, we extended the maximum length to 150, recognizing that restricting it to 50 tokens might truncate crucial target information occurring beyond this threshold. Given the complexity of this model, training necessitated a significant amount of time, prompting us to limit the number of epochs to 1 to expedite training. While we refrained from extensive experimentation to optimize hyperparameters, opting instead to adhere closely to the configuration used in the Fine-tuned BERT Model, this deliberate choice was made to ensure consistency and facilitate a more meaningful comparison.

Remarkably, this model exhibited a substantial improvement in accuracy compared to its predecessor.

3.2.4 Target-Dependent Fine-tuned multilingual BERT Model

The structure of this model mirrors that of the preceding one, with the sole variation being the utilization of a different pre-trained BERT model, namely 'bert-base-multilingual-cased'. Considering the multilingual nature of our dataset, we are intrigued to observe the potential enhancements in performance that may arise from employing a

BERT model trained on multilingual data.

4 Results & Discussion

The evaluation of model performances primarily relies on test accuracy and F1 scores. Test accuracy serves as a direct measure, representing the ratio of correctly classified instances to the total instances within the test set. Within sentiment classification, accuracy offers a broad assessment of the model's proficiency in accurately categorizing sentiments.

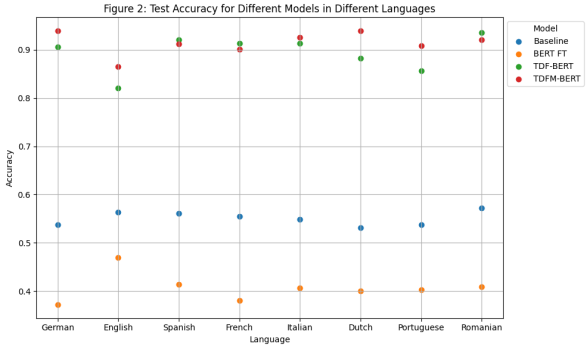
The F1 score, a composite metric derived from precision and recall, provides a balanced evaluation of classifier performance, particularly beneficial in scenarios with imbalanced datasets. While our dataset exhibits a predominantly balanced distribution, with neutral instances outnumbering negative and positive instances (as evidenced in Table 1), the F1 score remains pertinent. By accounting for both false positives and false negatives, it furnishes a comprehensive evaluation metric, ensuring robustness in performance assessment.

4.0.1 Test Accuracy

Table 2 delineates notable disparities in test accuracy across a diverse array of languages and model architectures. As delineated in Figure 2, it becomes conspicuous that target-dependent models (both TDF and TDFM) consistently exhibit significantly superior performance compared to the baseline and Bert FT models. This corroborates the notion that target-dependent methodologies indeed serve as potent tools for enhancing model efficacy.

Language/Model	Baseline	BERT FT	TDF-BERT	TDFM-BERT
German	0.538	0.372	0.906	0.939
English	0.563	0.469	0.821	0.865
Spanish	0.561	0.414	0.921	0.912
French	0.555	0.381	0.913	0.901
Italian	0.549	0.406	0.913	0.926
Dutch	0.531	0.4	0.882	0.939
Portuguese	0.538	0.403	0.857	0.908
Romanian	0.572	0.409	0.936	0.921

Table 2 Test Accuracy for Different Models in Different Languages



Furthermore, the BERT FT model consistently demonstrates diminished performance relative to the baseline model across various linguistic domains. This recurrent observation suggests that the BERT fine-tuning strategy may lack optimal suitability for this particular task, as it consistently falls short even when juxtaposed against simpler baseline models.

Upon conducting a comparative analysis between TDF and TDFM, discernible disparities in performance emerge across languages, with notable inconsistencies observed among disparate model configurations. While TDF demonstrates heightened performance in languages such as Spanish, French, and Romanian, TDFM outperforms in the remaining linguistic domains. This nuanced variability underscores the intricate nature of model performance across divergent linguistic landscapes. Notably, TDFM's efficacy appears comparatively subdued in Spanish, French, and Romanian, relative to other linguistic contexts. However, a noteworthy observation arises when TDF exhibits superior performance—the discernible performance gap between TDF and TDFM diminishes. Conversely, in instances where TDFM prevails, it attains heightened accuracy across a broader spectrum of languages, thus signifying its adaptability in multilingual settings despite nuanced performance variations contingent upon specific linguistic attributes and model specifications.

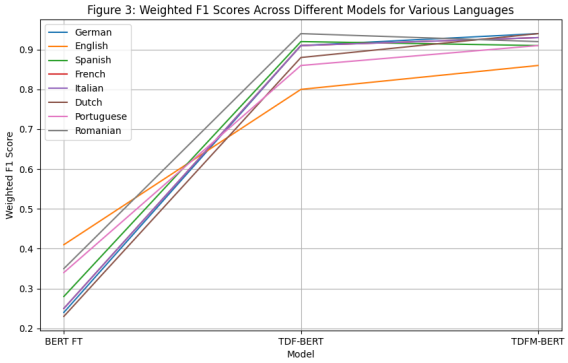
Another intriguing observation pertains to the unexpectedly diminished performance of target-dependent models in the English language domain. Both TDF and TDFM exhibit suboptimal performance metrics for English, a phenomenon that deviates from expectations given English's status as a high-resource language in the domain of natural language processing. Future investigations could aim to elucidate the underlying factors contributing to this observed phenomenon.

4.0.2 F1 Scores

Table 3 and Figure 3 illustrate the weighted F1 scores across different models for a variety of languages. The observed trends in F1 scores closely align with our findings regarding test accuracy. Across all languages, the target-dependent models (TDF and TDFM) consistently yield notably higher F1 scores compared to the BERT FT model, highlighting the efficacy of target-dependent architectures for the given task.

Language/Model	BERT FT	TDF-BERT	TDFM-BERT
German	0.24	0.91	0.94
English	0.41	0.8	0.86
Spanish	0.28	0.92	0.91
French	0.25	0.91	0.93
Italian	0.25	0.91	0.93
Dutch	0.23	0.88	0.94
Portuguese	0.34	0.86	0.91
Romanian	0.35	0.94	0.92

Table 3 Weighted F1 score Across Different Models for Various Languages



Although TDFM-BERT generally exhibits superior performance, there are instances where TDF-BERT achieves comparable or even superior F1 scores, such as in the case of Romanian. This suggests that the selection between TDF-BERT and TDFM-BERT may vary depending on the specific language and the requirements of the task at hand.

5 Conclusion

Our research contributes significantly to advancing the comprehension of multilingual sentiment analysis by harnessing the MAD-TSC dataset and conducting an extensive comparative examination of sentiment classification models. Utilizing evaluation metrics such as test accuracy and F1 scores, we elucidate the superior performance of target-dependent models, notably surpassing the effectiveness of the BERT fine-tuning approach. Furthermore, we affirm the efficacy of MBert in contrast to the base Bert model. These findings underscore the potency of multilingual models in addressing sentiment analysis tasks and emphasize the necessity of accounting for linguistic diversity in NLP research endeavors. Future investigations may delve deeper into the intricate performance variations observed across diverse languages and explore the underlying factors shaping model effectiveness in multilingual contexts.

References

[1] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June 2022. European Language Resources Association.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Evan Dufraisse, Adrian Popescu, Julien Tourille, Armelle Brun, and Jerome Deshayes. Mad-tsc: A multilingual aligned news dataset for target-dependent sentiment classification. 2023.

[4] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In Spandana Gella, Johannes Welbl, Marek Rei, Fabio Petroni, Patrick Lewis, Emma Strubell, Minjoon Seo, and Hannaneh Hajishirzi, editors, *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July 2020. Association for Computational Linguistics.

A MAD-TSC dataset

The MAD-TSC dataset example data:

English: For La Stampa, “it was clear from the encounter in Rome between Monti, Merkel, Rajoy and Hollande” on June 22 that the Council meeting, which gets underway this Thursday, “will be a tough and tricky test that will see European leaders try to establish and, who knows, bring in a revamped monetary union”:

Spanish: Para La Stampa, “desde el encuentro en Roma el 22 de junio entre Monti, Merkel, estaba claro" que el Consejo que se inicia este jueves “será la primera y dura prueba con la que los dirigentes europeos van a intentar fundar, y quizás lanzar, una nueva Unión Monetaria”:

German: Voor La Stampa werd het vorige week al tijdens de ontmoeting van Monti, Merkel, Hollande en Rajoy in Rome duidelijk dat de eurotop die vandaag in Brussel van start is gegaan, “de eerste lastige test is voor de Europese leiders omdat zij zullen trachten een nieuwe monetaire unie te bewerkstelligen en, wie weet, deze ook tot uitvoering te brengen”.

Italian: Per La Stampa, “era chiaro sin dal meeting tra Monti, Merkel, Hollande e Rajoy” del 22 giugno che il Consiglio che si inaugura questo giovedì “sarà la prima e dura prova attraverso la quale i dirigenti europei cercheranno di istituire e - chissà - forse lanciare una nuova unione monetaria”:

French: Pour La Stampa, “il était clair dès la rencontre romaine entre Monti, Merkel, Hollande et Rajoy” du 22 juin que le Conseil qui s’ouvre ce jeudi “constituera la première, et la plus difficile, des épreuves par lesquelles les dirigeants européens seront passés pour tenter de fonder et, qui sait, de lancer, une nouvelle Union monétaire” :

Portuguese: Para La Stampa, "é evidente, desde o encontro entre Monti, Merkel, Rajoy e Hollande", em 22 de junho, que o Conselho que começa nesta quinta-feira "será a primeira e a mais dura prova em que os dirigentes europeus vão procurar estabelecer e, quem sabe, introduzir uma nova União Monetária":

Dutch: Für La Stampa ist es seit dem Treffen von Monti, Merkel, Hollande und Rajoy in Rom am 22. Juni klar, dass der am Donnerstag beginnende Gipfel „die erste schwere Prüfung sein wird, mit der die europäischen Staats- und Regierungschefs versuchen werden, eine neue Währungsunion zu gründen und, wer weiß, zu starten“:

Romanian: Pentru La Stampa, "era clar încă de la întâlnirea romană dintre Monti, Merkel, Hollande şi Rajoy", din 22 iunie, că acest Consiliu care a început astăzi "va fi prima şi cea mai grea încercare prin care conducătorii europeni vor încerca să fondeze, şi, cine ştie, să lanseze, o nouă uniune monetară":