International Journal of Service and Knowledge Management
International Institute of Applied Informatics
2024, Vol. 8, No. 2, IJSKM850

The Influence of Linguistic Attribute Differences in Multilingual Datasets on Sarcasm Detection

Linshuo Yang [*], Daisuke Ikeda [*]

Abstract

Presently, social media is crucial for sentiment analysis tasks with machine learning. However, the presence of sarcasm presents a challenge to this task by concealing the true intent of a text. Consequently, there has been a surge in research on automatic sarcasm detection. Furthermore, studying with sarcasm detection with multilingual datasets is becoming indispensable, because it can sovle data scarcity problem for low-resource languages and can also reduce the cost of training models for different languages. Past research has largely overlooked the influence of language diversity within training datasets on model performance. This study assumes that linguistic differences may influence sarcastic expressions and employs two datasets: English- Arabic dataset, which belongs to the same category of morphological typology, and English-Chinese dataset which belongs to different categories. Subsequently, models were trained with BERT, BERT-BiLSTM, and BERT-RCNN architectures. Finally, results were compared using two English test datasets with different patterns. The outcomes revealed superior training results for English-Arabic in contrast to English-Chinese, signifying the influence of morphological typology. In addition, BiLSTM and RCNN architectures can enhance the performance of multilingual sarcasm detection models. And the RCNN structure appears to be beneficial for detecting sarcasm in different patterns.

Keywords: machine learning, morphological typology, Multilingual BERT, multilingual sarcasm detection

1 Introduction

The progress of deep learning has significantly advanced natural language processing technology, particularly within the subfield of sentiment analysis. Sentiment analysis involves the classification of texts into positive, negative, or neutral categories, and its applications span a wide range of fields, including customer service, marketing, and social media.

[*] Kyushu University, Fukuoka, Japan

However, sentiment analysis encounters several challenges, including the complexity and subjectivity nature of human emotions, limited availability of labeled training data, and the presence of sarcasm. Sarcasm is recognized as a linguistic form employed to convey the opposite of the literal meaning, poses a unique challenge. Sarcasm is often used humorously to express dissatisfaction or conflict while conveying emotions indirectly, and it has become prevalent in social media where users may find it challenging to express themselves directly. For example, a user wrote these words on the SNS: "love it when Jonny makes me use earphones to watch something on my phone so he can play fifa". From a literal perspective, this text appears to express the user's fondness and admiration for Jonny's behavior. However, it's easy to see that the person is actually trying to convey his/her shock and disapproval of Jonny's rude behavior. Additionally, using the same way to express sarcastic may serves as a means to establish connections and form groups based on shared values and interests. However, detecting sarcasm in real-life interactions relies on factors like tone, expression, and context. On social media, sarcasm is not always clearly identifiable from the text alone, leading to potential misunderstandings. As sarcasm can significantly alter the intended meaning and emotion of a message, accurately detecting sarcasm has become a critical task of sentiment analysis.

Previous research [1, 2] has mainly concentrated on sarcasm detection in English. However, as the use of social media extends globally, online communication is carried out in diverse languages, introducing increased diversity and complexity of online speech content. Given that sarcasm's linguistic attributes and cultural context vary across languages, the task of automatically detecting sarcasm becomes even more challenging. Therefore, there is a crucial demand for research on multilingual sarcasm detection. Moreover, the shortage of datasets for sarcasm detection is currently a significant challenge. Multilingual sarcasm detection can improve this situation by augmenting the amount of training data, thereby enhancing the accuracy of sentiment analysis for social media communication across various languages. Additionally, research on multilingual sarcasm detection hold the potential to clarify the linguistic and cultural characteristics of sarcasm, laying the groundwork for cross-cultural communication applications in the field of artificial intelligence.

Prior multilingual sarcasm detection research has primarily focused on classification using lexical or syntactic approaches. Recently, deep learning models have also emerged as viable options. However, previous studies have excessively emphasized model development while underestimating the influence of multilingual data. In an effort to explore the influence of language types within training datasets, this study starts from morphological typology and hypothesizes that models trained on data from linguistically similar languages perform better in their own language type. Notably, to the best of the authors' knowledge, there is a dearth of prior research investigating the impact of morphological typology on multilingual sarcasm detection. To investigate this hypothesis, this study combines a Chinese dataset and an Arabic dataset respectively with an identical English dataset, thus forming two mixed-language datasets. This choice stems from the fact that Arabic and English both belong to the inflectional language category, whereas Chinese belongs to the analytic language category.

Subsequently, these two dataset sets are utilized to train three different models constructed based on the Multilingual-BERT model to substantiate the hypothesis. Initially, training was conducted using Multilingual BERT alone as the base line. After that, the BERT-BiLSTM model was employed to investigate the results of deepening the learning of overall text features. In this model, after extracting features using Multilingual BERT, bidirectional LSTM structures were used for learning. To further investigate the impact of local textual features on the results, an additional experiment was carried out utilizing BERT-RCNN model in addition to the results in our previous article [3]. This model involved feature extraction using BERT, followed by

BiLSTM and CNN structures to learn both global and local patterns. As a result, BERT-BiLSTM and BERT-RCNN structure was found to be generally more effective than BERT-only model. And RCNN structure is proved to be helpful in cross-pattern sarcasm detection.

The remainder of this paper is as follows: Section 2 briefly introduces the development of sarcasm detection techniques. Section 3 provides an overview of the datasets used in this study and the preprocessing applied to these datasets for experimentation. Section 4 elaborates on the methodology employed in this research and presents the results in detail. Section 4 provides a detailed explanation of the methods and process of the experiment, and then presents the results. Finally Section 5 concludes this research and outlines the future directions.

2 Related Work

This section provides an overview of previous research methods in sarcasm detection and their limitations.

After Tepperman et al. [4] initially proposed the concept of sarcasm detection in speech, it has been attracting attention in the field of natural language processing. Sarcasm detection is generally treated as a classification problem, and early research primarily focused on utilizing machine learning models to identify features of sarcasm. Riloff et al. [5] and Joshi et al. [6] emphasized the presence of affirmative language use in describing negative situations as a mark of sarcasm, and they developed models based on this concept. Subsequently, Reyes et al. [7] were the first to employ this contradiction as a means to detect sarcasm in social media. Ghosh et al. [8] used the contradiction within individual tweets as features and employed Support Vector Machines (SVM) for sarcasm detection. As research progressed, the role of context was also recognized as a valuable feature in sarcasm detection. Wallace et al. [9] demonstrated that introducing previous statements made by the commenter and the responses to the comments could enhance sarcasm detection performance, particularly on platforms like Reddit.

Deep learning meethods can capture complicated sentence features [10], significantly reducing the effort required for feature identification. Consequently, research employing deep learning methods for sarcasm detection has gained widespred attention in recent years. Recurrent neural networks (RNNs) have been proved effective in understanding the sequential relationships between words in a sentence, and an attention mechanism based on Long Short-Term Memory (LSTM) has been proposed [10]. Notably, Kumar et al. [1] achieved state-of-the-art performance at that time by employing a combination of word embeddings and Convolutional Neural Networks (CNNs), an approach referred to as CNN-LSTM. Recently, pre-trained language models such as BERT [11] and ERNIE [12] have been emerged as the new state-of-the-art in natural language processing. Khatri et al. [2] conducted a comparative analysis between feature-based methods and fine-tuning methods utilizing pre-trained models for sarcasm detection in English tweets. Their findings highlighted the superiority of the latter approach. Lai et al. [13] introduced a model structure named Recurrent Convolutional Neural Network (RCNN), which can simultaneously capture global sequence features and local features. Their experimental results achieved state-of-the-art performance at that time, demonstrating the effectiveness of this structure.

In recent years, the detection of sarcasm in independent sentences has gained significant attention, not only in English but also in various other languages, such as Italian [14] and Japanese [15]. Furthermore, there has been active research in multilingual sarcasm detection to meet the global demand for natural language processing tasks. Abbasi et al. [16] proposed an entropy-weighted algorithm and applied it to a mixed-language text dataset containing both English

and Arabic. Joshi et al. [17] introduced an LSTM-based architecture capable of learning sub-word-level representations, particularly for sentiment analysis within a dataset that combines Hindi and English. Jain et al. [18] presented a model structure based on softmax bidirectional LSTM and CNN for the Hindi-English mixed-language dataset, ultimately enhancing sarcasm detection performance. Furthermore, Han et al. [19] developed a method involving fine-tuning pre-trained models such as ERNIE-M and DeBERTa for sarcasm detection in both English and Arabic datasets prepared in SemEval-2022 Task 6[1] [20]. Their approach achieved outstanding performance in this context.

In this paper, the authors also investigated multilingual sarcasm detection utilizing the same datasets as in [19]. Notably, prior research has predominantly concentrated on understanding the inherent characteristics of "sarcasm" and devising efficient methods for extracting these characteristics, often overlooking the potential influence of the dataset itself. This paper focus on the language types within multilingual datasets, and particularly investigates the morphological typological differences between Arabic and Chinese and explores how these differences influence sarcasm detection methods when employing deep learning techniques.

3 Datasets

This section introduces the datasets utilized in the experiments conducted in this study. Considering the quality of the datasets and the ease of acquisition, we chose datasets in English, Arabic, and Chinese for the experiments. Among these languages, English is selected as the foundational dataset due to its larger quantity and higher quality. The training data in English and Arabic are from the dataset provided by SemEval-2022 Task 6 [20]. Meanwhile, the training data in Chinese is sourced from the"Open Chinese Internet Sarcasm Corpus" publicly released by Zhu [21]. Given the objectives of our experiments, we chose to use English testing datasets, is the same as the language of the foundational dataset. For the English test datasets, SemEval-2022 Task 6 provides a portion, and additional test data is obtained from the "News Headlines Dataset For Sarcasm Detection" accessible on Kaggle[2] [22, 23]. Below, we provide a comprehensive overview of each of these datasets.

3.1 SemEval Datasets

SemEval-2022 Task 6, proposed by Ibrahim Abu Farha and his team, focuses on the development of sarcasm detection systems utilizing datasets in both English and Arabic. Notably, both the training and test data for this task have been meticulously annotated by the team. In many existing annotation methods for text sarcasm detection, the labeling process relies on predefined criteria to identify and annotate sarcastic content. For example, texts that contain specific tags, such as "#sarcasm" or "#irony", are often marked as instances of sarcasm.

However, this annotation method may introduce noisy data for several reasons:

(1) Tags may not necessarily indicate sarcasm but might simply represent parts of sentence structure, as exemplified by phrases like "#sarcasm is all around!.

(2) The assumption that a specific tag consistently indicates sarcasm or that a particular social

[1] https://sites.google.com/view/semeval2022-isarcasmeval
[2] https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

media account consistently posts sarcastic statements can be incorrect.

(3) Categorizing texts that do not meet the predefined criteria as non-sarcastic may result in misclassification.

To deal with the aforementioned issues, Abu Farha et al. proposed a novel data collection method. This approach seeks to mitigate external errors by directly involving the speakers themselves in providing labels for the data.

Data collection method:

(1) English: For English text, they recruited the assistance of English native speakers via the Prolific Academic platform[3]. This platform enables researchers to engage participants online. The participants were requested to supply one sarcastic tweet and three non-sarcastic tweets that they had previously posted.

(2) Arabic: Native Arabic speakers were recruited with the "Appen" platform[4], which facilitates data sourcing and preparation. These speakers were tasked with generating sarcastic sentences. Non-sarcastic examples were from the ArSarcasm-v2 dataset [24].

The following examples illustrate sarcastic and non-sarcastic English tweets from the dataset:

- sarcastic: The only thing I got from college is a caffeine addiction
- non-sarcastic: Exams have finally finished!! I can be happy again #woooooo

In English training dataset, there are 867 sarcastic examples and 2601 non-sarcastic ones. And for Arabic training dataset, the number of sarcastic examples is 745 and 2357 for non-sarcastic ones.

3.2 News Headlines Dataset For Sarcasm Detection

Misra et al. [22, 23] compiled a noise-free dataset by gathering news headlines from two distinct news websites. Non-sarcastic news headlines were obtained from Huff- Post, a reputable and serious news website. Sarcastic data was acquired from a website called "The Onion", which imitates HuffPost by crafting sarcastic versions of current news topics. Since non-sarcastic content is never actually published on "The Onion", this dataset is free from noise. This dataset has garnered widespread attention and has been downloaded more than 35,000 times, making it a favored choice in numerous sarcasm detection studies due to its exceptional performance.

[3] https://www.prolific.co/
[4] https://appen.com/

The following examples show sarcastic and non-sarcastic headlines from the dataset:

- sarcastic: romney volunteers going door-to-door to let obama supporters know president's dead
- non-sarcastic: Dem Rep. Totally Nails Why Congress Is Falling Short On Gender, Racial Equality

3.3 Open Chinese Internet Sarcasm Corpus

Existing sarcasm detection datasets are primarily available in English, and high-quality Chinese datasets are notably scarce. Consequently, Zhu took the initiative to create a Chinese dataset in 2022, employing the following methodology:

(1) Collecting a large amount of sarcastic text is difficult due to the relatively low proportion of such content. To augment the corpus size without compromising quality, the existing corpus data is utilized.

(2) To expedite the collection and annotation process, a strategy known as selecting a "high-density source" is employed. This means choosing sources where either sarcastic or non-sarcastic materials are more likely to appear, reducing the time required for dataset creation and labeling.

The sarcastic data in this corpus originates from two sources. The first source is a corpus constructed by Tang and Chen [25], where tags were appended to each sentence element. Zhu converted these tags into binary classification annotations for sarcasm detection. Furthermore, since the initial corpus contained numerous short sentences, Zhu augmented the dataset by gathering data from one of China's largest social platforms, Weibo[5]. This data was accurately sourced from an account named "Yangcong Gushihui" (Onion The Storyteller), which is renowned for its high density of sarcastic news and comments, similar to the style of "The Onion". Zhu manually checked this data, eliminating irrelevant information such as advertisements, and specifically selected longer sarcastic messages for inclusion. Conversely, the non-sarcastic data was collected from official Weibo accounts in China known for publishing serious news content. There are 1000 positive examples and 1000 negative examples collected.

3.4 Data Augmentation

As mentioned earlier, the training datasets for SemEval has a relatively small amount of sarcastic data, whether in English or Arabic. And the datasets are highly unbalanced. Therefore, in this study, data augmentation was employed utilizing ChatGPT[6]. In this study, the ChatGPT API was utilized to imitate all the sarcastic texts in both the English and Arabic datasets. In other words, for each positive example, a text with the same meaning but different wording was generated, effectively doubling the dataset size. The final number of sarcastic examples is 1734 for English and 1490 for Arabic. Compared to the previous data augmentation method of directly copying the original text to increase the quantity, the approach used in this study not only increased the

[5] https://weibo.com/
[6] https://chat.openai.com

dataset size but also avoided the occurrence of repetitive patterns. The generated text was then checked by the authors by their own knowledge and translation tools to ensure that the texts indeed remain sarcastic.

3.5 Preprocessing of the Dataset

Across all the datasets previously described, a standard preprocessing step was applied, involving the removal of URLs and mentions.

As discussed in Section 3.1, the SemEval English and Arabic datasets are inherently unbalanced, which can potentially adversely affect model training results. Moreover, since the primary objective of this study is to explore the influence of morphological typology on multilingual sarcasm detection, two datasets were utilized for two distinct language pairs: English-Chinese and English-Arabic, for training the model. As the Chinese dataset comprises a balanced dataset with 1000 positive and 1000 negative examples, and to ensure that unnecessary disparities between the two mixed datasets are eliminated, an equivalent number of positive and negative examples (1000 for each) were randomly sampled for Arabic data. And the negative examples in the English dataset were also randomly sampled to match the number of positive examples.

4 Experiment

In this section, we will explain our experiments. We leveraged the pre-trained Multilingual BERT model as a feature extraction tool. Our approach involved training the model on two multilingual datasets, with subsequent comparison of the results using the same English test dataset. Furthermore, additional experiments by adding a Bidirectional LSTM (BiLSTM) structure or RCNN structure to BERT. Then, the outcomes of these model structures were discussed.

4.1 Experimental Procedure

For this study, the pre-training model "bert-base-multilingual-cased" available on Huggingface[7] was employed. Huggingface is an AI community that champions opensource contributions. In contrast to the standard BERT model, Multilingual BERT is trained on diverse language datasets, enabling multiple languages to be mapped into the same vector space. Given the necessity to handle both characters and lowercase alphabets in this study, the "cased" version of the model was utilized.

Table 1: hyperparameters

parameters	values
batch size	32
epochs	5
LSTM hidden size	256
LSTM dropout	0.1
optimizer	AdamW
learning rate	4e-6

[7] https://huggingface.co/

The hyperparameters of the models employed in the experiments in following sections are shown in Table 1.

In this study, the datasets were subdivided into ten equal portions while preserving the proportion of positive and negative examples. For each training iteration, two of these portions were designated as validation data, while the remaining eight parts were utilized for model training. This approach enabled the utilization of the entire datasets, efficiently utilized the data to its fullest extent. In the experimental phase, for each model structure, the model was first trained on the English-Chinese and English-Arabic datasets. Subsequently, it was tested using both the SemEval English test dataset and the dataset described in Section 3.2.

4.2 Model Structure

In our experiments, we explored three distinct model structures. Initially, we employed the Multilingual BERT model for embedding purposes and then directed the data through a fully connected layer for subsequent processing. Subsequently, we introduced a BiLSTM layer between the Multilingual BERT embedding and the fully connected layer. This modification aimed to investigate the impact of considering both forward and backward sequences on the model's performance and results. Finally, we conducted an additional experiment where we replaced the previously mentioned BiLSTM layer with an RCNN structure. In this setup, the vectors processed by the BiLSTM layer were concatenated with the feature vectors extracted by BERT. Then, the vectors passed CNN and max-pooling structures, followed by classification through a fully connected layer.

4.2.1 Structure with BERT only

The structure of the model is shown in Figure 1.

First, sentence data is processed by the pre-trained model Multilingual BERT, which extracts features and converts words into 768-dimensional word embeddings. As the embedding of the first token ([CLS] token) of the processed sentence by BERT encapsulates information about the entire sentence, it is employed as the sentence's embedding. Next, this embedding is passed through a fully connected layer, transforming it into a two-dimensional output. Finally, the position number of the larger dimension in this two-dimensional output is utilized as the label for the detection result.

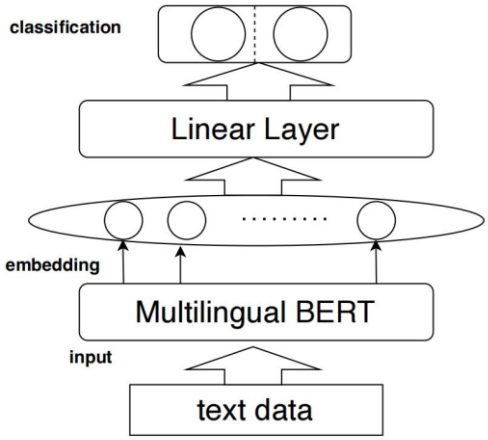

Figure 1: data embedded with BERT and then go through fully connected layer to get detection result

4.2.2 Structure with BERT-BiLSTM

LSTM is a type of layer employed in neural networks that solves the vanishing gradient problem and enhances the capabilities of traditional Recurrent Neural Networks (RNNs). Bidirectional LSTM (BiLSTM), on the other hand, is a variation of the LSTM network. It consists of two LSTM layers operating in opposite directions, which allows it to take into account both forward and backward sequences. Leveraging this bidirectional feature, it becomes possible to integrate information not only from the preceding sentence but also from the following sentence. In this study, we utilized the network depicted in Figure 2 to emphasize the significance of relationships in both directions and to explore the effects of morphological typological classification differences.

In this network, similar to Section 4.2.1, the data is initially transformed into a 768-dimensional embedding by Multilingual BERT. This obtained embedding representation is subsequently fed into the BiLSTM layer. In a BiLSTM, there exist two LSTM networks that process the input sequence in both forward and backward directions. The outputs from the final hidden layer of these two LSTM networks represent the combined features from the forward and backward directions, respectively. These two hidden outputs are then concatenated, resulting in a feature vector that takes into account the word order within the sentence. Given that the LSTM hidden size is set to 256, this feature vector becomes of dimension 2*256. Subsequently, this

feature vector passes through two fully connected layers to yield a two-dimensional output. The network structure is illustrated in Figure 2.

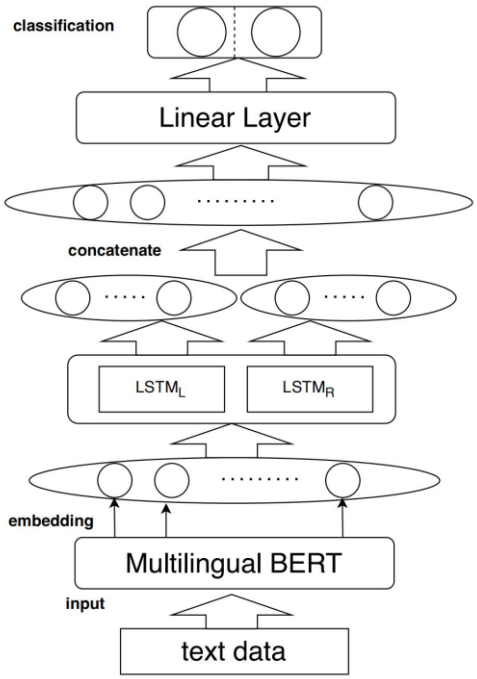

Figure 2: data embedded with BERT and processed by LSTM then go through fully connected layer to get detection result

4.2.3 Structure with BERT-RCNN

RCNN is a structure that combines RNN and CNN, and it can fully leverage the advantages of both networks. As mentioned earlier, RNN helps capture the global information of the text sequence, while CNN is good at learning local text features. The network structure is depicted in Figure 3.

In this network, similar to the first two networks, the Multilingual BERT is responsible for transforming input data into 768-dimensional feature vectors. These vectors then pass through a BiLSTM layer to learn global features in both forward and backward directions. The output of the BiLSTM's hidden layer is vectors of size 2*256. These vectors would be concatenated with the feature vectors extracted by BERT and then processed through a CNN to learn local text features. Finally, a fully connected layer is used for binary classification.

4.3 Result

The models trained as outlined in Section 4.1 is subjected to be tested using the SemEval English test data and the News Headlines Dataset for Sarcasm Detection. This section will present the results of these tests and provide a discussion of the outcomes.

Accuracy and F1-score are used as evaluation criteria for the test results. Accuracy is a simple performance metric used to evaluate a classification model. It measures the proportion of correctly classified samples out of the total number of samples. The formula for accuracy is:

$$Accuracy = \frac{Number\ o\ f\ Correctly\ Classified\ Samples}{Total\ Number\ of\ Samples} \tag{1}$$

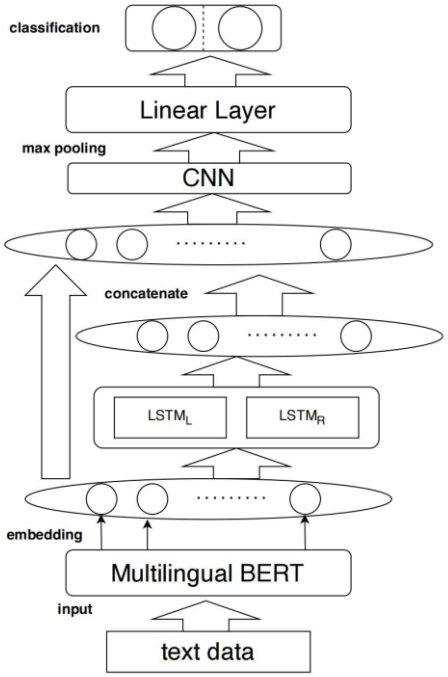

Figure 3: data embedded with BERT and processed by LSTM and CNN and max-pooling and finally go through fully connected layer to get detection result

On the other hand, F1-score is a more comprehensive performance metric that combines a model's precision (measures how many of the predicted positive samples are actually true positives) and recall (measures how many of the true positive samples were successfully identified by the model), making it particularly useful for dealing with unbalanced classes. The formula for calculating the F1-score is as follows (where TP represents True Positives, FP represents False Positives, and FN represents False Negatives):

$$\begin{cases} Precision = \dfrac{TP}{TP + FP} \\ Recall = \dfrac{TP}{TP + FN} \\ F1 - score = \dfrac{2 * Precision * Recall}{Precision + Recall} \end{cases} \tag{2}$$

The results on the SemEval test dataset are as shown in Table 2. Because the seminal test dataset is highly unbalanced with 200 positive examples and 1200 negative examples, the evaluation primarily relies on the F1-score instead of just focusing on accuracy. It can be observed that, apart from the BERT-only model, the other two models, which learned more text information, achieved better results on the En-Ar dataset. Furthermore, the introduction of the

BiLSTM layer indeed improved the model's performance, and this improvement was more pronounced on the En-Ar dataset. In other words, when focusing on the multilingual case and considering both the forward and backward sequences, the training dataset with Arabic, which falls into the same classification as English, outperforms the Chinese case on English test data.

As a matter of fact, in the original paper summarizing the SemEval competition by Abu et al. [20], the metric used to evaluate the performance of various models on the English test set was also the F1-score. Abu et al. also established two baselines, baseline-bert and baseline-svm. Table 3 lists some of the competition results and rankings at that time. Teams stce [26], X-PUDU [19], and TUG-CIC [27] got the top three positions with a significant lead. The En-Ar BERT-LSTM and En-Ar BERTRCNN models, which performed well in this study, both outperformed the baseline. Furthermore, they ranked between Team underfined [28] at the 14th position and Team CS-UM6P [29] at the 15th position in the competition results at that time. This indicates that the data augmentation and model training in this experiment was executed effectively.

Table 2: The results of SemEval dataset

Model	F1-score	Accuracy
En-Cn BERT-only	0.344	0.774
En-Cn BERT-BiLSTM	0.365	0.774
En-Cn BERT-RCNN	0.313	0.768
En-Ar BERT-only	0.303	0.734
En-Ar BERT-BiLSTM	0.382	0.781
En-Ar BERT-RCNN	0.375	0.793

Table 3: Compare the results of this study with some of the SemEval-2022 task 6

Ranking	Model/Team	F1-score
1	stce	0.605
2	X-PuDu	0.569
3	TUG-CIC	0.530
14	underfined	0.383
-	En-Ar BERT-LSTM	0.382
-	En-Ar BERT-RCNN	0.375
15	CS-UM6P	0.371
19	baseline-bert	0.348
27	baseline-svm	0.275

The results from the News Headlines test dataset are presented in Table 4. As the News Headlines test dataset is balanced, accuracy can be the criterion. As shown in the table, compared to the results of the SemEval dataset, the F1-score and accuracy of the News Headlines dataset have significantly decreased. This can be attributed to the fact that the SemEval data is derived from daily conversations, while News Headlines consist of highly formal article headlines. These two datasets have very different patterns, which can explain why the introduction of BiLSTM for learning overall text patterns did not yield better test results compared to the BERT-only model in this context.

Remarkably, when Arabic is introduced, both the F1-score and accuracy consistently outperform the results obtained with Chinese, irrespective of the model structure. It's worth noting

that the Chinese dataset, as mentioned in Section 3.3, contains texts that imitate The Onion, which is the source of sarcastic examples in the News Headlines dataset. From this perspective, the English-Chinese pair may appear to be closer in terms of test patterns. However, the results demonstrate that pairs involving English and Arabic, which belong to the same morphological typology, yield even better results. This outcome serves as strong evidence that the morphological typology of languages indeed exerts an influence on the results of multilingual sarcasm detection.

Table 4: The results of News Headlines dataset

Model	F1-score	Accuracy
En-Cn BERT-only	0.113	0.515
En-Cn BERT-BiLSTM	0.089	0.524
En-Cn BERT-RCNN	0.271	0.538
En-Ar BERT-only	0.180	0.528
En-Ar BERT-BiLSTM	0.130	0.532
En-Ar BERT-RCNN	0.314	0.549

Furthermore, we can observe that although the results on the News Headlines test dataset are not particularly impressive for all models, BERT-RCNN stands out comparatively. This indicates that while the introduction of the BiLSTM structure may make the model susceptible to text patterns, the ability of CNN to learn local text features can effectively mitigate this influence, thus enhancing the model's cross-pattern capabilities.

5 Conclusion

This study focused on the role of morphological typology in multilingual sarcasm detection, and conducted comprehensive data processing and experiments to explore this aspect. The outcomes of these experiments are compelling. When utilizing the same training model and evaluating on the same English dataset, the model trained on the English-Arabic pair, which falls within the same category, consistently outperformed the model trained on the English-Chinese pair, which belongs to different categories. This provides a clear indication that the typological classification of languages indeed exerts a significant influence on the results of multilingual sarcasm detection. Furthermore, the cross-pattern capabilities of the RCNN model have also been validated.

This conclusion leads to two directions. First, in current multilingual sarcasm detection research, if the target language for detection is determined, it is crucial to use language data that belongs to the same category as the target language to construct the training dataset. This approach is expected to enhance the model's performance in detecting sarcasm effectively. Second, it is equally important to investigate new data processing methods and training techniques aimed at mitigating the challenges associated with mixing different types of language data.

Regarding future directions of work, first of all, given the scarcity of sarcasm detection datasets, particularly for languages other than English, researchers should explore innovative strategies to overcome these challenges. Furthermore, while this study tested the hypothesis on English, Chinese, and Arabic datasets, future research should aim to conduct experiments on larger datasets encompassing a broader range of languages. This expansion will contribute to a more comprehensive understanding of the influence of morphological typology on multilingual sarcasm detection and lead to more effective models for a variety of languages.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23K28149.

References

[1] Lakshya Kumar, Arpan Somani, and Pushpak Bhattacharyya. “Having 2 hours to write a paper is fun!”: Detecting Sarcasm in Numerical Portions of Text. *arXiv preprint arXiv:1709.01950*, 2017.

[2] Akshay Khatri et al. Sarcasm detection in tweets with BERT and GloVe embeddings. *arXiv preprint arXiv:2006.11512*, 2020.

[3] Linshuo Yang and Daisuke Ikeda. The Impact of Language Properties in Multilingual Datasets on Sarcasm Detection. *the 14th International Congress on Advanced Applied Informatics* (IIAI-AAI), *the 16th International Conference on E-Service and Knowledge Management* (ESKM 2023), pages 1–6, 2023.

[4] Joseph Tepperman, David Traum, and Shrikanth Narayanan. “YEAH RIGHT”: SARCASM RECOGNITION FOR SPOKEN DIALOGUE SYSTEMS. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, 2006.

[5] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714, 2013.

[6] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing Context Incongruity for Sarcasm Detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, 2015.

[7] Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in Twitter. *Language resources and evaluation*, 47:239–268, 2013.

[8] Aniruddha Ghosh and Tony Veale. Fracking Sarcasm using Neural Network. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 161–169, 2016.

[9] Byron C Wallace, Eugene Charniak, et al. Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, 2015.

[10] Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. Reasoning with Sarcasm by Reading In-between. *arXiv preprint arXiv:1805.02856*, 2018.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv preprint arXiv:1904.09223*, 2019.

[13] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent Convolutional Neural Networks for Text Classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[14] Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2892–2899, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[15] Satoshi Hiai and Kazutaka Shimada. Sarcasm Detection Using Features Based on Indicator and Roles. In *Recent Advances on Soft Computing and Data Mining: Proceedings of the Third International Conference on Soft Computing and Data Mining (SCDM 2018), Johor, Malaysia, February 06-07, 2018*, pages 418–428. Springer, 2018.

[16] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM transactions on information systems (TOIS)*, 26(3):1–34, 2008.

[17] Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491, 2016.

[18] Deepak Jain, Akshi Kumar, and Geetanjali Garg. Sarcasm detection in mashup language using soft-attention based bi-directional LSTM and feature-rich CNN. *Applied Soft Computing*, 91:106198, 2020.

[19] Yaqian Han, Yekun Chai, Shuohuan Wang, Yu Sun, Hongyi Huang, Guanghao Chen, Yitong Xu, and Yang Yang. X-PuDu at Semeval-2022 Task 6: Multilingual Learning for English and Arabic Sarcasm Detection. *arXiv preprint arXiv:2211.16883*, 2022.

[20] Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States, July 2022. Association for Computational Linguistics.

[21] Yizhang Zhu. Open Chinese Internet Sarcasm Corpus Construction: An Approach. *Frontiers in Computing and Intelligent Systems*, 2(1):7–9, 2022.

[22] Rishabh Misra and Prahal Arora. Sarcasm Detection using Hybrid Neural Network. *arXiv preprint arXiv:1908.07414*, 2019.

[23] Rishabh Misra and Jigyasa Grover. *Sculpting Data for ML: The first act of Machine Learning.* Jan. 2021.

[24] Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021.

[25] Yi-jie Tang and Hsin-Hsi Chen. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1269–1278, 2014.

[26] Mengfei Yuan, Zhou Mengyuan, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. stce at SemEval-2022 Task 6: Sarcasm Detection in English Tweets. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval- 2022)*, pages 820–826, 2022.

[27] Jason Angel, Segun Aroyehun, and Alexander Gelbukh. TUG-CIC at SemEval- 2021 Task 6: Two-stage Fine-tuning for Intended Sarcasm Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval- 2022)*, pages 951–955, 2022.

[28] Xiyang Du, Dou Hu, Jin Zhi, Lianxin Jiang, and Xiaofeng Shi. PALI-NLP at SemEval-2022 Task 6: iSarcasmEval-Fine-tuning the Pre-trained Model for Detecting Intended Sarcasm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 815–819, 2022.

[29] Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, Abderrahman Skiredj, and Ismail Berrada. CS-UM6P at SemEval-2022 Task 6: Transformerbased Models for Intended Sarcasm Detection in English and Arabic. *arXiv preprint arXiv:2206.08415*, 2022.