X-PuDu at SemEval-2022 Task 6: Multilingual Learning for English and Arabic Sarcasm Detection

Yaqian Han, Yekun Chai, Shuohuan Wang, Yu Sun — Baidu Inc., Beijing, China

Hongyi Huang, Guanghao Chen, Yitong Xu, Yang Yang — Shanghai Pudong Development Bank

Abstract

Detecting sarcasm and verbal irony from people's subjective statements is crucial to understanding their intended meanings and real sentiments and positions in social scenarios. This paper describes the X-PuDu system that participated in SemEval-2022 Task 6, iSarcasmEval-Intended Sarcasm Detection in English and Arabic, which aims at detecting intended sarcasm in various settings of natural language understanding. Our solution finetunes pre-trained language models, such as ERNIE-M and DeBERTa, under the multilingual settings to recognize the irony from Arabic and English texts. Our system ranked second out of 43, and ninth out of 32 in Task A: one-sentence detection in English and Arabic; fifth out of 22 in Task B: binary multi-label classification in English; first out of 16, and fifth out of 13 in Task C: sentence-pair detection in English and Arabic.

Introduction

Sarcasm is the use of language that typically signifies the opposite to mock or convey contempt. As a narrow research field in natural language processing (NLP), sarcasm detection is a particular case in the spectrum of sentiment analysis, with important implications for a slew of NLP tasks, such as sentiment analysis, opinion mining, author profiling, and harassment detection. In the textual data, these tonal and gestural clues like heaving tonal stress and rolling of the eyes are missing, making it more difficult for machines. For sarcastic texts, the authors also rephrase them into non-sarcastic ones. Then, linguistic experts further checked the scathing pieces and labeled them into sub-categories of sarcasm defined by Leggitt[1] : sarcasm, irony, satire, understatement, overstatement, and rhetorical question. Our method employed various multilingual or mono-lingual pre-trained language models, such as ERNIE-M[2] and DeBERTa[3] to address each component of this task, with a bunch of fine-tuning and ensemble techniques.

Methods

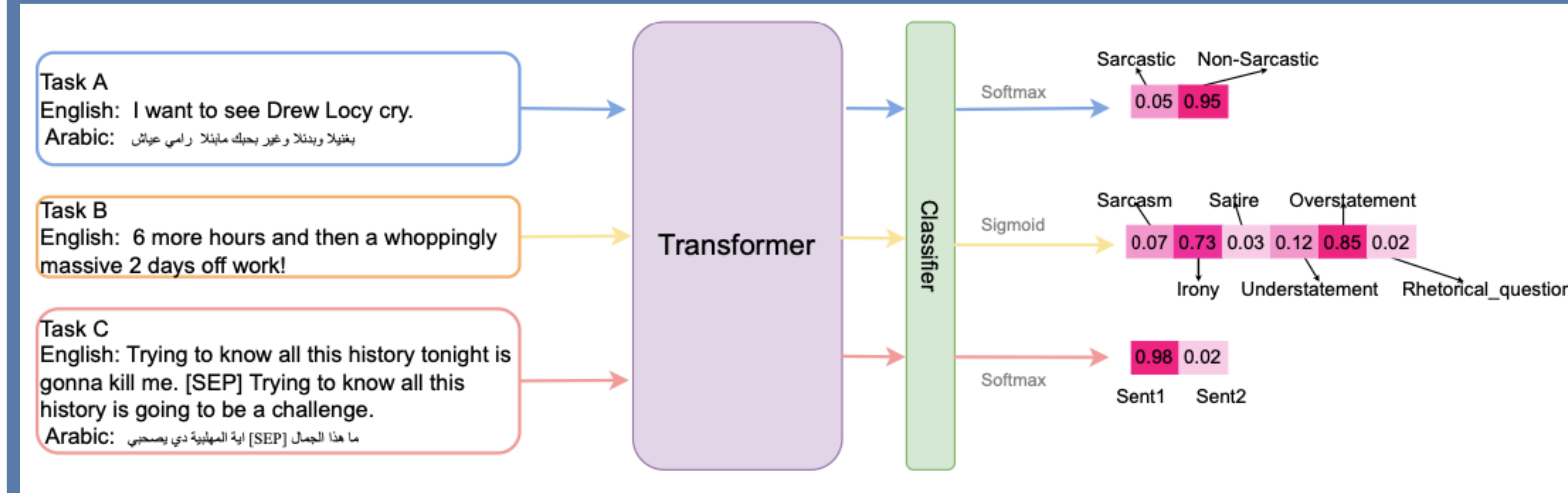

Figure 1: Fine-tuning pre-trained models on the *iSarcasmEval* data.

• **Pre-trained Language Models** We adopt pretrain-then-finetune paradigm for better leveraging the performance of large-scale pre-trained models. As illustrated in Figure 1, for all tasks, we utilize pre-trained models to extract the input representations, followed by a fully-connected feed-forward layer and a softmax/sigmoid activation after the [CLS] token for prediction. For sub-task A and B that input samples only contain one sentence, we directly fine-tune the pre-trained Transformers. For sub-tasks with two sentences, i.e., sub-task C, we employ the multi-layer pre-trained Transformer blocks as the cross-encoder by concatenating sentence pairs and separating them with a [SEP] token.

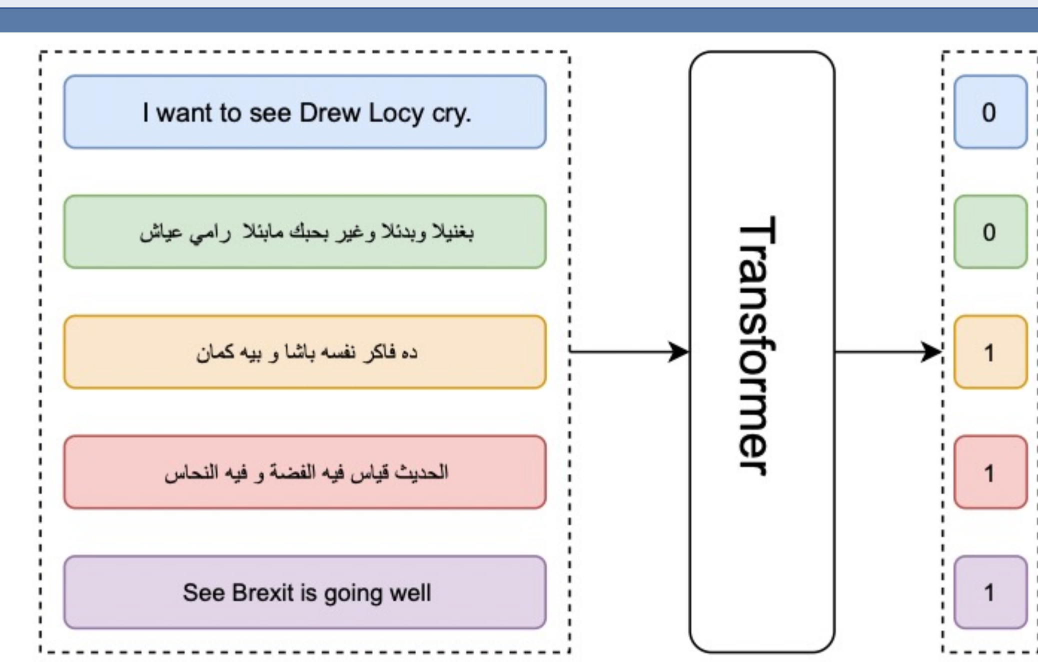

Figure 2: Multilingual learning on Task A. "0/1" indicate the non-sarcastic and sarcastic class.

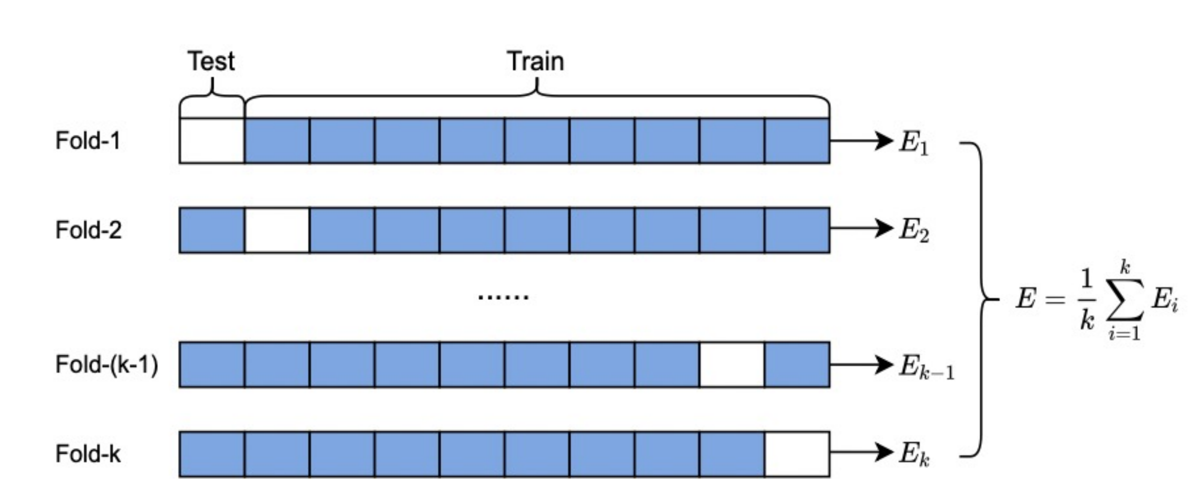

Figure 3: Illustration of ensemble learning. E_i indicates the prediction of the i-th model on the test set.

• **Multilingual Learning** We adopt multilingual learning method by simultaneously fine-tuning the pre-trained models on both Arabic and English training data based on multilingual pre-trained models, i.e., ERNIE-M. As shown in Figure 2, we combine the one-sentence binary sarcasm detection subtasks in English and Arabic together and fine-tune the multilingual pre-trained models in one forward pass.

• **Ensemble Learning** We adopt 10-fold cross-validation for training as a way to improve the robustness of our model, as shown in Figure 3.

Results

Task	Lang	ERNIE-M (multilingual)	ERNIE-M (monolingual)	DeBERTa	Rank
Task A	en	36.75	38.46	**56.91**(∗)	2/43
	ar	40.36	**41.87**(∗)	-	9/32
Task B	en	N/A	-	**7.99**(∗)	5/22
Task C	en	82.50	75.00	**87.00** (∗)	1/16
	ar	**90.50**	84.00(∗)	-	5/13

Table 1: Official test-set performance under various experimental settings. The "ERNIE-M (multilingual)" column indicates the performance of multilingual learning in Task A and C. Scores with asterisk indicate final submitted results. The official evaluation metrics for Task A,B,C are F1-score, macro F1-score, and accuracy, respectively.

Conclusions

We present our system that participated in SemEval Task 6 and employ the multilingual learning method to train the English and Arabic tasks jointly. We empirically find that it confers benefits in specific scenarios and outranks the monolingual pre-trained models on Arabic tasks. However, we do not adopt other Arabic-specific pre-trained models, which is also worth comparing. In the future, it is a promising direction to explore different sarcasm detection approaches under multilingual settings.

References

[1] John S. Leggitt and Raymond W. Gibbs. 2000. Emotional reactions to verbal irony. Discourse Processes, 29(1):1–24.

[2] Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora.

[3] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient disentangled embedding sharing.