CS-UM6P at SemEval-2022 Task 6: Transformer-based Models for Intended Sarcasm Detection in English and Arabic

Abdelkader El Mahdaouy Abdellah El Mekki Kabil Essefar[†]
Abderrahman Skiredj[†] Ismail Berrada[†]
School of Computer Science, Mohammed VI Polytechnic University, Morocco
abdelkader.elmahdaouy@um6p.ma, abdellah.elmekki@um6p.ma
{firstname.lastname}@um6p.ma[†]

Abstract

Sarcasm is a form of figurative language where the intended meaning of a sentence differs from its literal meaning. This poses a serious challenge to several Natural Language Processing (NLP) applications such as Sentiment Analysis, Opinion Mining, and Author Profiling. In this paper, we present our participating system to the intended sarcasm detection task in English and Arabic languages. Our system[1] consists of three deep learning-based models leveraging two existing pre-trained language models for Arabic and English. We have participated in all sub-tasks. Our official submissions achieve the best performance on sub-task A for Arabic language and rank second in sub-task B. For sub-task C, our system is ranked 7th and 11th on Arabic and English datasets, respectively.

1 Introduction

Sarcasm is an important aspect of human natural language. It is characterized by the occurrence of a discrepancy between the intended and the literal meaning of utterance (Wilson, 2006). The prevalence of this phenomenon may jeopardize the performance of many NLP applications, such as Sentiment Analysis, Opinion Mining, and Emotion Detection, among others (Maynard and Greenwood, 2014; Rosenthal et al., 2014; Van Hee et al., 2018). Indeed, sarcasm detection has been the subject of many systematic investigation, where several shared tasks have been organized and a number of datasets have been introduced (Van Hee et al., 2018; Ghanem et al., 2019; Ghosh et al., 2020; Abu Farha et al., 2021). The existing datasets are either labeled by a human annotator or using distant supervision signals such as the presence of a set of predefined hashtags (Oprea and Magdy, 2020). However, these labeling methods might be sub-optimal for intended sarcasm detection as the author's sarcastic intent may differ from an annotator's perceived meaning (Oprea and Magdy, 2019, 2020). Besides, most existing research works have focused on English language (Van Hee et al., 2018; Ghosh et al., 2020), while few studies have been introduced for other languages such as Arabic (Ghanem et al., 2019; Abu Farha et al., 2021).

To overcome the aforementioned limitations, Abu Farha et al. (2022) have organized the iSarcasmEval shared task for intended sarcasm detection in English and Arabic languages. In contrast to previous research work, the introduced dataset, for intended sarcasm detection, is labeled by the authors themselves. The authors are then asked to provide non-sarcastic rephrases that covey the same intended meaning of their sarcastic texts. Further, the iSarcasmEval's organizers have relied on linguistic experts to categorize sarcastic texts into sarcasm, irony, satire, understatement, overstatement, and rhetorical questions (Leggitt and Gibbs, 2000).

In this paper, we present our participating system to iSarcasmEval shared task (Abu Farha et al., 2022). Our system rely on three transformer-based deep learning models. In all our models, we use existing Pre-trained Language Model (PLM) to encode the input text and apply a single attention layer to the contextualized word embedding of PLM (Barbieri et al., 2021; Abdul-Mageed et al., 2021). For all our models, we use the same classifier architecture composed of one hidden layer and one classification layer. The classifier is fed with the concatenation of the pooled output of the PLM as well as the attention layer output. We model the sub-task A as a binary classification (first model) and as a multi-class classification (second model). Motivated by the small size of the datasets and similarly to GAN-BERT architecture (Croce et al., 2020), the third model uses a conditional generator that tries to generate fake samples that are similar to the PLM's embedding of the real input data. The

[1] The source code of our system is available at https://github.com/AbdelkaderMH/iSarcasmEval

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 844 - 850
July 14-15, 2022 ©2022 Association for Computational Linguistics

discriminator of the third model is trained to discriminate between fake and real samples as well as classify the real ones correctly. For sub-task B, we only employ the GAN based model (third model), while for sub-task C, we utilize all models trained on Task A and we compare the probabilities of the sarcastic class of the left and the right text. We train our models using several loss functions, including the focal loss (Lin et al., 2017). Besides, for sub-task A, we train our models with and without the non-sarcastic rephrase texts.

For the official submissions, we employ hard voting ensemble of our trained models. Our system achieve promising results as we rank 1st, 15th, 2nd, 7th, and 11th on sub-task A AR, sub-task A EN, sub-task B, sub-task C AR, and sub-task C EN, respectively.

2 Background

2.1 Task description

The organizers of iSarcasmEval shared task have provided training data and testing data for intended sarcasm detection in English and Arabic languages (Abu Farha et al., 2022). The datasets are collected from Twitter and labeled for intended sarcasm detection by the authors of the tweets. For English, the training data consists of 3,468 samples out of which 867 samples are sarcastic. For Arabic, the training data contains 3,102 samples, where 745 samples are sarcastic. The organizers also provide the non-sarcastic rephrase of sarcastic texts for both languages and the dialect of the given Arabic tweets. The shared task consists of the flowing sub-tasks:

- **Sub-task A** is a binary classification task, where the aim is to determine if a tweet is sarcastic or not. This sub-task consists of two sub-tasks A EN and A AR. The test data contains 1,400 for each language.

- **Sub-task B** is a multi-label classification task, where the aim is to assign a given tweet into the sarcasm, irony, satire, understatement, overstatement, and rhetorical question categories of ironic speech (Leggitt and Gibbs, 2000). This task is provided for English language only and the test data contains 1,400 samples.

- **Sub-task C** aims to identify the sarcastic tweet and the non-sarcastic rephrase given two texts that convey the same meaning. This sub-task consists of two sub-tasks C EN and C AR. The test sets of both languages consist of 200 samples.

2.2 Related work

In recent years, there has been a growing number of research works focusing on fine-tuning the existing PLMs on NLP tasks. These PLMs are based on the transformer architecture and are trained using self-supervised learning objectives such as Masked Language Modeling (MLM) amongst others (Devlin et al., 2019). Several multilingual and monolingual PLM variants are introduced (Devlin et al., 2019; Conneau et al., 2020; Antoun et al., 2020). For domain-specific data, domain adaptive fine-tuning of existing PLMs using MLM or domain adaptation have been shown to improve the performance of NLP applications (Rietzler et al., 2020; Barbieri et al., 2021; El Mekki et al., 2021a). Nevertheless, when the domain-specific data is sufficiently large, these transformers can be trained from scratch (Abdul-Mageed et al., 2021; Inoue et al., 2021).

For sarcasm detection, several research studies have been introduced based on fine-tuning the existing PLMs for English and Arabic languages (Ghanem et al., 2019; Ghosh et al., 2020; Abu Farha et al., 2021). El Mahdaouy et al. (2021) have shown that incorporating attention layers on top of the contextualized word embedding of the PLM improves the performance of multi-task and single-task learning models for both sarcasm detection and sentiment analysis in Arabic. The main idea consists of classifying the input text based on the concatenation of the PLM's pooled output and the output of the attention layer. This Architecture has yielded promising results on other tasks such as detecting and rating humor, lexical complexity prediction, and fine-grained Arabic dialect identification (Essefar et al., 2021; El Mamoun et al., 2021; El Mekki et al., 2021b).

Although transformer-based architectures have shown state-of-the-art performance on many NLP tasks, their task-specific fine-tuning requires a reasonable amount of labeled data. Nevertheless, in real-world applications, one may not always have enough labeled training data. Motivated by the performance of Semi-Supervised Generative Adversarial Networks (SS-GAN) in computer vision (Odena, 2016), Croce et al. (2020) have introduced GAN-BERT. The latter extends BERT fine-tuning procedure by training a generator and a discriminator. The generator is trained to generate fake samples embeddings that are similar to the real input text embeddings, whereas, the discriminator is trained to detect fake examples and categorize the real text inputs.

3 System overview

The submitted system to the iSarcasmEval shared task relies on three transformer-based models for intended sarcasm detection. In order to encode the input text, we utilize the Twitter-XLM-Roberta-base (Barbieri et al., 2021) and MARBERT (Abdul-Mageed et al., 2021) for English and Arabic texts, respectively. The former is a variant of XLM-RoBERTa PLM that is adapted to Twitter data using MLM objective, while the latter is a variant of BERT PLM that is pre-trained from scratch on Arabic tweet corpora. In the following subsection, we describe the components of our system.

3.1 Preprocessing

The tweet preprocessing component spaces out emojis and substitutes user's mention and URL with their special tokens of the PLM's tokenizer. For Twitter-XLM-Roberta-base, URLs and user's mentions are replaced by 'http' and '@user'. For MARBERT, they are replaced by 'user' and 'url' special tokens. To leverage the dialect information for Arabic data, we replace the dialect string with its full Arabic name and pass the input text to MARBERT's tokenizer as follows:

- [SEP] dialect [SEP] preprocessed text [SEP]

3.2 Deep Learning Models

Our three deep learning models are described as follows:

- **Model 1** consists of a transformer encoder, one attention layer, and a classifier. Following the work of (El Mahdaouy et al., 2021), we apply attention to the contextualized word embedding of the encoder. The classifier is composed of one hidden layer and one classification layer for binary classification. The classifier is fed with the concatenation of the PLM's pooled output and the attention layer's output.

- **Model 2** is similar is to **Model 1** and the task is modeled as multi-class classification problem. In other words, the classification layer of this model consists of two hidden units. Figure 1 illustrates the overall architecture of Model 1 and Model 2.

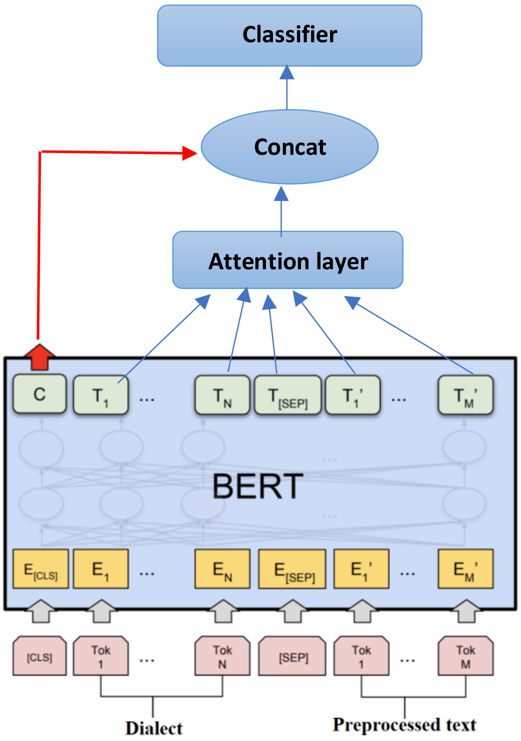

Figure 1: The overall architecture of Model 1 and Model 2.

- **Model 3** is similar to GAN-BERT model (Croce et al., 2020), whereas, we employ a conditional generator that generates fake embeddings from a random noise and the class category. The model 3 consits of three components, a BERT-based encoder with an extra attention layer on top of the contextualized word embedding, a generator, and a discriminator. The encoder represents the input sentences using the CLS token embedding and the output of the attention layer. The generator is trained to fool the discriminator by generating fake embedding representations that are similar to the input text embedding. It consists of two hidden layers and one output layer. Each hidden layer is flowed by a dropout layer and the relu activation layer. The discriminator is trained to discriminate between fake examples and real ones and to classify real input texts into sarcastic and non-sarcastic labels. The discriminator is also composed of two hidden layers and one classification layer. Similarly to the generator, the hidden layers are followed by one dropout layer and the relu activation layer.

3.3 Training objectives

For training our models, we utilizes several loss functions, including the Focal Loss (Lin et al., 2017). The aim is to assess the performance of the following training objectives under class imbalance:

- For **Model 1**, we investigate the Binary Cross-Entropy loss (BCE), the Weighted Binary Cross-Entropy loss (W. BCE), and the Binary Focal Loss (BFL). For the W. BCE loss, the positive class weight is set to:

$$\frac{batch\_size - positive\_count}{positive\_count + \epsilon}$$

- For **Model 2 and 3**, we investigate the Cross-Entropy loss, the Weighted Cross-Entropy loss (W. CE), and the Focal Loss (FL). For the W. CE, the positive and the negative class weights are computed as follows:

$$\begin{cases} pos\_weight = \frac{batch\_size - positive\_count}{positive\_count + \epsilon} \\ neg\_weight = \frac{batch\_size - negative\_count}{negative\_count + \epsilon} \end{cases}$$

4 Experimental setup

All our models are implemented using the PyTorch[2] framework and the open-source Transformers[3] libraries. Experiments are conducted on a PowerEdge R740 Server having 44 cores Intel Xeon Gold 6152 2.1GHz, a RAM of 384 GB, and a single Nvidia Tesla V100 with 16GB of RAM. 20% of the training set is used for the model validation. All our models are trained using Adam optimizer with a linear learning rate scheduler. Based on our preliminary results, obtained on the validation set, the learning rate, the number of epochs, and the batch size are fixed to 1×10^{-5}, 10, and 16 respectively. For the focal loss, the hyper-parameters γ and α (the weight of the negative class) are set to 2 and 0.8 respectively. All models are evaluated using the Accuracy as well as the macro averaged Precision, Recall, and F1 measures. Besides, we train our models with and without rephrase texts for sub-task A. For sub-task B, our models are trained on sarcastic tweets.

5 Results

In this section, we present the obtained results of our models as well as our official submissions. It is worth mentioning that for sub-task C, we employ the trained models on sub-task A and we use their output probabilities to discriminate between the sarcastic text and the non-sarcastic rephrase. Besides, for our official submissions, we use the hard vote ensemble of our trained models. For each loss function, the best performance obtained is highlighted in **bold** font, while the overall best performance for each task is highlighted in ***bold-italic*** font.

5.1 Sub-task A

Table 1 summarizes the obtained results. The results show that training the models on tweets as well as the non-sarcastic rephrases improve the performance for both languages, especially Arabic where an important performance increment is yielded. Moreover, the performance of the evaluated models depends on the employed loss function. Although Model 1 and Model 2 are simple, they achieve better results than the GAN-based model (Model 3). The best performances on sub-task A are obtained using Model 2 in conjunction with the FL loss (0.6217) and Model 1 in conjunction with BCE loss (0.3833) for Arabic and English respectively. In our official submission, ensembling the models that are trained with and without the rephrase data harms the results. Our submitted system achieves the best performance on Arabic and ranked 15th on the English.

5.2 Sub-task B

Table 2 presents our obtained results for sub-task B. Since we use the sarcastic tweets only for training, we only train the Model 3. The results show that the best performance is obtained using the BCE loss. The FL and W. BCE loss functions have not improved the results. This might be explained by the fact that we did not tune the hyper-parameters α and $gamma$ of FL loss. Besides, in the W. BCE, the positive classes are assigned larger importance weights in comparison to the negative ones (see Section 3.3). Our official submission yields the second-best results on this sub-task.

5.3 Sub-task C

Table 3 summarizes the obtained results on sub-task C test sets of Arabic and English languages. In accordance with the results of sub-task A, the

[2] https://pytorch.org/

[3] https://huggingface.co/transformers/

Table 1: The obtained results on the test set of sub-task A for both Arabic and English. For our non official submissions, we report the macro F-1 score of the sarcastic class only.

	Sub-Task A Arabic						**Sub-Task A English**					
	Tweet only			Tweet + rephrase			Tweet only			Tweet + rephrase		
	BCE/CE	BFL/FL	W. BCE/CE	BCE	BFL/FL	W. BCE/CE	BCE/CE	BFL/FL	W. BCE/CE	BCE/CE	BFL/FL	W. BCE/CE
Model 1	0.5253	**0.5621**	0.4565	**0.6135**	0.5787	**0.5793**	0.3485	**0.3605**	0.3421	***0.3833***	0.3313	**0.3714**
Model 2	0.5307	0.4481	0.4892	0.5949	***0.6217***	0.5505	**0.3574**	0.3375	0.3144	0.3770	**0.3619**	0.3090
Model 3	**0.56**	0.5271	**0.4937**	0.5339	0.5488	0.5345	0.3470	0.3448	**0.3478**	0.3517	0.3517	0.3557
	Official Submission						**Official Submission**					
Ensembling	F-1 sarcastic	F-score	Precision	Recall	Accuracy	**Rank**	F-1 sarcastic	F-score	Precision	Recall	Accuracy	**Rank**
	0.5632	0.7188	0.6948	0.8362	0.8050	**1**	0.3713	0.6171	0.6058	0.6458	0.7750	**15**

Table 2: The obtained results on the test set of sub-task B.

		F-1 Macro	F-1 sarcasm	F-1 irony	F-1 satire	F-1 understatement	F-1 overstatement	F-1 rhetorical question
Model 3	BCE	**0.0924**	**0.2331**	0.1676	0.0530	0	0	**0.1008**
	BFL	0.0877	0.2298	**0.1733**	0.025	0	0	0.0983
	W. BCE	0.0681	0.2302	0.0705	**0.0581**	0	0	0.0501
Ensembling	**Rank**	**Official Submission**						
	2	0.0875	0.2314	0.1622	0.0392	0	0	0.0923

Table 3: The obtained results on the test set of sub-task C for both Arabic and English. For our non-official submissions, we report the accuracy score only.

	Sub-Task C Arabic						**Sub-Task C English**					
	Tweet only			Tweet + rephrase			Tweet only			Tweet + rephrase		
	BCE/CE	BFL/FL	W. BCE/CE	BCE	FL	W. BCE/CE	BCE/CE	BFL/FL	W. BCE/CE	BCE/CE	BFL/FL	W. BCE/CE
Model 1	**0.68**	**0.69**	0.61	0.815	**0.82**	0.83	**0.69**	**0.68**	**0.67**	0.695	**0.7**	**0.685**
Model 2	0.65	0.575	0.59	**0.835**	0.815	***0.85***	0.655	**0.68**	0.65	0.685	0.67	0.655
Model 3	0.67	0.68	**0.685**	**0.835**	0.815	0.835	0.655	**0.68**	0.625	***0.71***	0.685	**0.685**
	Official Submission						**Official Submission**					
Ensembling	Accuracy	F-1 Score	**Rank**				Accuracy	F-1 Score	**Rank**			
	0.7800	0.7688	**7**				0.6950	0.6481	**11**			

models that are trained on the tweets and the non-sarcastic rephrases yield better performances. The best-obtained accuracy scores are 0.85 and 0.71 in comparison to 0.78 and 0.69 obtained by our official submission for Arabic and English respectively. Hence, ensembling the models that are trained with and without the rephrase data harms the performance of our official submission. Our official submission is ranked 7th and 11th on Arabic and English respectively.

6 Conclusion

In this paper, we present our participating system in the iSarcasmEval shared task for intended sarcasm detection in English and Arabic. Our system relies on three deep learning-based models that leverage two existing pre-trained language models for Arabic and English. We participate in all sub-tasks, investigate several training objectives, and we study the impact of including non-sarcastic rephrase in the training data. The results show that ensembling models that are trained with and without rephrases have a negative impact on the official results. Our official submissions achieve the best performance on sub-task A for the Arabic language and rank in the second position on sub-task B.

Acknowledgments

Experiments presented in this paper were carried out using the supercomputer Toubkal, supported by Mohammed VI Polytechnic University (https://www.um6p.ma), and facilities of simlab-cluster HPC & IA platform.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

7088–7105, Online. Association for Computational Linguistics.

Ibrahim Abu Farha, Silviu Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2021. A Multilingual Language Model Toolkit for Twitter. In *arXiv preprint arXiv:2104.12250*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021. Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 334–339, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nabil El Mamoun, Abdelkader El Mahdaouy, Abdellah El Mekki, Kabil Essefar, and Ismail Berrada. 2021. CS-UM6P at SemEval-2021 task 1: A deep learning model-based pre-trained transformer encoder for lexical complexity. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 585–589, Online. Association for Computational Linguistics.

Abdellah El Mekki, Abdelkader El Mahdaouy, Ismail Berrada, and Ahmed Khoumsi. 2021a. Domain adaptation for Arabic cross-domain and cross-dialect sentiment analysis from contextualized word embedding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2837, Online. Association for Computational Linguistics.

Abdellah El Mekki, Abdelkader El Mahdaouy, Kabil Essefar, Nabil El Mamoun, Ismail Berrada, and Ahmed Khoumsi. 2021b. BERT-based multi-task model for country and province level MSA and dialectal Arabic identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 271–275, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kabil Essefar, Abdellah El Mekki, Abdelkader El Mahdaouy, Nabil El Mamoun, and Ismail Berrada. 2021. CS-UM6P at SemEval-2021 task 7: Deep multi-task learning model for detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1135–1140, Online. Association for Computational Linguistics.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. Idat at fire2019: Overview of the track on irony detection in arabic tweets. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 10–13, New York, NY, USA. Association for Computing Machinery.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. 2020. A report on the 2020 sarcasm detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 1–11, Online. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *CoRR*, abs/1708.02002.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4238–4243, Reykjavik, Iceland. European Language Resources Association (ELRA).

Augustus Odena. 2016. Semi-supervised learning with generative adversarial networks.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743. Language in Mind: A Tribute to Neil Smith on the Occasion of his Retirement.