stce at SemEval-2022 Task 6: Sarcasm Detection in English Tweets

Mengfei Yuan and **Mengyuan Zhou** and **Lianxin Jiang** and **Yang Mo** and **Xiaofeng Shi**
PALI Inc.
Shenzhen, China
{YUANMENGFEI854, ZHOUMENGYUAN425, JIANGLIANXIN769,
MOYANG853, SHIXIAOFENG309}@pingan.com.cn

Abstract

This paper describes the systematic approach applied in "SemEval-2022 Task 6 (iSarcasmEval) : Intended Sarcasm Detection in English and Arabic". In particular, we illustrate the proposed system in detail for SubTask-A about determining a given text as sarcastic or non-sarcastic in English. We start with the training data from the officially released data and then experiment with different combinations of public datasets to improve the model generalization. Additional experiments conducted on the task demonstrate our strategies are effective in completing the task. Different transformer-based language models, as well as some popular plug-and-play proirs, are mixed into our system to enhance the model's robustness. Furthermore, statistical and lexical-based text features are mined to improve the accuracy of the sarcasm detection. Our final submission achieves an F1-score for the sarcastic class of 0.6052 on the official test set (the top 1 of the 43 teams in "SubTask-A-English" on the leaderboard).

1 Introduction

Sarcasm is a sophisticated communication technique to express emotions, attitudes, feelings, and evaluations. Sarcastic and ironic texts typically do not contain words with negative polarity, hostile attitudes, or offensive in their literal sense, but rather express the contradiction or opposite of the literal meanings (Filik et al., 2016; Van Hee et al., 2018; Reyes and Rosso, 2014; Verma et al., 2021). Sarcasm detection can be considered a particular sentiment analysis task, applied to detect texts that are intended to use some exaggeration, understatement, or rhetoric content to express criticism or praise for people or events. Many researchers have conducted different deep learning methods (Poria et al., 2016; Kumar et al., 2020; Zhang et al., 2019), traditional machine learning method (Buschmeier et al., 2014; Hernández-Farías et al., 2015; Yaghoobian et al., 2021), and big data approaches (Bharti et al., 2016; Sarsam et al., 2020; Ortega-Bueno et al., 2019) to improve the accuracy of irony or sarcasm auto-detection.

SemEval-2022 Task 6 (iSarcasmEval) is a sarcasm detection task (Abu Farha et al., 2022). The standard training dataset includes 3468 English tweets and 3102 Arabic tweets. The English training dataset provides 862 sarcastic tweets along with their non-sarcastic rephrases, while the Arabic datasets provides 745 sarcastic samples. For the English dataset, each sarcastic tweet is also labeled as a fine-grained multi-class and multi-label ironic tag such as satire, understatement, overstatement, and rhetorical questions.

SubTask-A is a binary classification task to predict whether a given tweet is sarcastic or not. Table 1 shows one sarcastic tweet and its non-sarcastic version, and one non-sarcastic tweet from the released dataset. We could notice that the raw tweets are pretty noisy and contain user information, URLs, hashtags, etc. Some of the sarcastic tweets also contain sarcastic-related words such as "irony" or "sarcastic" in their hashtags. Many non-sarcastic tweets include confused, unfriendly words or denial attitudes.

In this paper, we demonstrate the following contributions: 1) The discrepancy in prediction performance using different transformer-based language models; 2) The improvement of adding the public dataset and mining effective text features; 3) The enhancement obtained by incorporating various constrative learning loss functions; 4) Model generalization is improved by incorporating the multi-sample dropout layer into the output of pre-trained language models. On Subtask-A, our system achieves an F1 score for the sarcasm category of 0.6052 and a Macro F1 score of 0.7675.

2 System Overview

The final submitted result is a contribution from various classification models using the voting mech-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 820 - 826

Sarcastic tweet	@PFTompkins Her family should definitely not seek mental health guidance.
Sarcastic tweet rephrase	They should seek guidance.
Non-sarcastic example	I wonder if it's too late for me to re-enroll in University and relive it all just one last time.

Table 1: Sarcastic and non-sarcastic tweet examples

anism. The outcome is a fusion of 15 predictions trained using different transformer-based language models, external datasets, text features, and deep learning-based techniques such as constrastive loss, adversarial training, multi-sample dropout, etc.

The proposed model is trained with a 5-fold cross-validation with a randomly distributed seed. The basic classification model is structured with a multi-sample dropout layer after the pooling layer of the pretrained model. RoBERTa-large, XLM-RoBERTa-large and DeBERTa-v3-large are alternatively adopted in these 15 models. Models are trained using the AdamW optimizer with a learning rate of 1e-05 in the fast gradient method. Four dropout layers with a rate of 0.4 are picked in our multi-sample dropout module. When we make models based on DeBERTa-v3-large, we change the dropout rate to 0.2, which has been suggested by previous studies (He et al., 2020b, 2021).

The cross-entropy loss (L_{Xent}) is used as the classification loss, and the additional XNET loss (L_{NTXent}) with a temperature of 0.2 is selected as the metric learning loss in our system. Equation 1 shows the combination of two types of loss. The weight parameter (w) used to balance the combination of multiple losses is set as 0.1.

$$Loss = (1 - w)L_{Xent} + wL_{NTXent} \quad (1)$$

Additionally, three external datasets are added to the official SemEval-2022 data in the proposed models for further training. Text features are directly concatenated to the training texts in some proposed models as well. The featuring mining, text preprocessing, and the voting method are described below. The scheme of data preparation, training, and prediction processes is demonstrated in Figure 1.

2.1 Pre-processor

The raw English tweets in the official training data contain many noises such as misleading hashtags, usernames, website links, and emojis in different formats. We detected and replaced usernames and links with special tokens. Additionally, we extended some common English abbreviations, such as U, idk, omg, sry, etc. to full-spelled words to keep the whole dataset in the same phase. However, we intend to keep words with unusual capitalization, wrong spelling, and repeated punctuation in raw tweets since people sometimes prefer to express exaggeration and emphasis in this way.

2.2 External Data

Besides the officially released data, we trained the model with three public datasets. The description and the source for additional data are illustrated as below.

(1) The iSarcasm [1] is a public dataset of English tweets. Each tweet is labeled as either sarcastic or non-sarcastic. Each sarcastic tweet is labeled with a fine-grained ironic label as well. We obtained 2279 non-sarcastic and 563 sarcastic tweets using the tweet API (Oprea and Magdy, 2020).

(2) The Multi-modal Sarcasm data [2] contains 33,859 images with descriptions where the sarcastic and non-sarcastic text are uniformly distributed. We used the text information only to train our classification model. In reality, we randomly took 10,000 texts out of the full dataset as the additional dataset. This strategy can speed up the training process and avoid the bias from a large number of additional datasets as well.

(3) The dataset released by SemEval-2018 task3 [3] is also considered in our training dataset (Van Hee et al., 2018). The data format and the task description are pretty similar to our task. This dataset contains 3800 tweets with uniform sarcastic/non-sarcastic labels.

[1] https://github.com/silviu-oprea/iSarcasm

[2] https://github.com/headacheboy/data-of-multimodal-sarcasm-detection

[3] https://competitions.codalab.org/competitions/17468

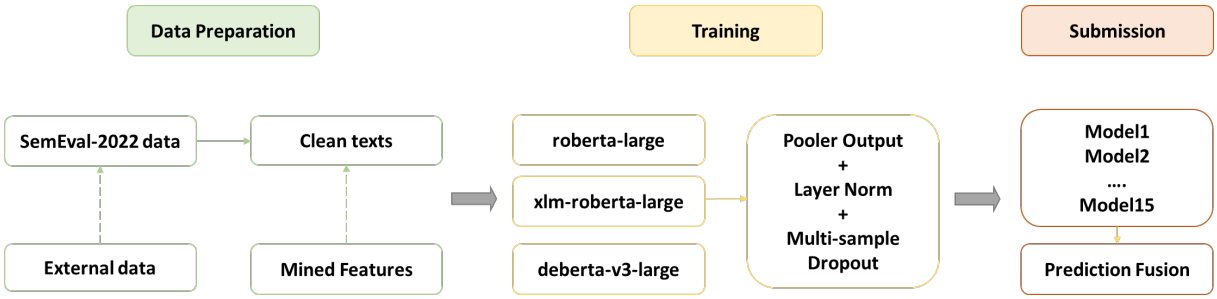

Figure 1: Task experimental progress

2.3 Language Models

We have adopted the RoBERTa-large (Liu et al., 2019) and DeBERTa-v3-large (He et al., 2020b, 2021) as the pretrained models from Hugging Face. We also applied the XLM-RoBERTa-large (Conneau et al., 2019) as a pretrained model for the dataset which includes the Arabic tweets during training.

2.4 Feature Mining

Moreover, we mined different types of statistical and lexical-based features that were previously applied in irony detection. Additional text features can improve the detection of sarcasm in many related tasks (Hernández-Farías et al., 2015; Yaghoobian et al., 2021). All the text features are simply added to the preprocessed tweets using the splitting token "$< /s > < /s >$". Figure 2 demonstrates how we concatenate different features into the original text.

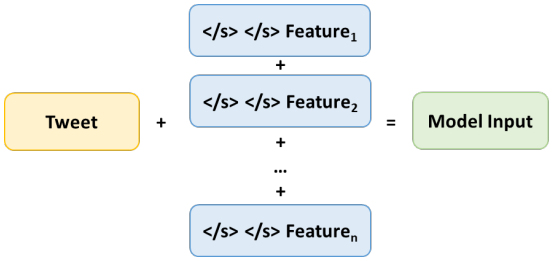

Figure 2: Text and features concatenation

(1) Emoji is a prominent multi-model feature that indicates human emotions after analyzing large amounts of tweet data [4]. On social media, emojis with few characters can easily turn common text into humorous, sarcastic, or ironic expressions.

(2) Parts-of-speech (POS) information is applied as an important feature as well. It is worth mentioning that we mined the POS-based features from the sarcastic tweets and their own rephrases in the official SemEval-2022 dataset. A list of adjectives and adverb words is generated by comparing the differences between the sarcastic tweets and their rephrased versions. For example, some words like "really", "never", "actually", etc. can be considered a symbol of sarcasm and express some contradictory and criticized attitudes.

(3) We also notice that some misspelled words (e.g., "so"->"soooo", "love"->"looove", "sure"->"sureeee") and capitalized words (not located at the beginning of a sentence) can sometimes exaggerate the emotional expression. Those words are detected and added to the tweets as additional text features.

(4) Transitional words and words with negative polarity are also considered as two potential sarcastic features (Tayal et al., 2014). Transitional words can express opposition or contradiction or indicate different meanings in a same sentence, such as "on the other hand" and "neverthelsess" [5]. The polarity of words or lexicon sometimes helps to identify the level of praise or criticism of a text. For the polarity-based feature, we adopted the AFINN dataset [6] (a list of words labelled with a polarity valence) as a reference.

2.5 Ensemble

The final submitted result is fused, utilizing the voting-based mechanism, with the predictions of 15 pretrained models. Voting from different models can usually hinder obvious mis-classifications from a single model (Ruta and Gabrys, 2005; Zhang et al., 2014). Hard voting and soft voting are two classical voting methods in classification tasks.

[4] https://github.com/MathieuCliche/Sarcasm_detector

[5] http://www.csun.edu/~hcpas003/transwords.html;https://wordcounter.net/blog/2016/07/19/101889_transition-words.html

[6] http://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt

Hard voting directly fuses different ensembles by picking the highest number of votes. The amount of sarcasm and non-sarcasm votes from different models directly determines the label of a test sample. Soft voting combines all ensembles by adding the probabilities of each prediction and picking the prediction with the highest probability summation. The predicted label would be sarcasm when the mean probability of the sarcasm category is greater than our selected threshold of 0.5. We mixed the hard and soft voting methods for the final submitted prediction. The hard voting method is adopted when the difference between the amount of sarcasm and non-sarcasm is greater than 2. Otherwise, the soft voting method is adopted.

3 Experimental Setup

Four major improvements in F1 score are given by adding the public dataset, multi-sample dropout layer, text features, and tuning the parameters in the contrastive loss function. In this paper, we evaluate different modules and tricks that are adopted in our proposed model to show their effect on the Semeval-2022 official dataset. Table 2 shows the F1 scores for the competition blind test set based on different pretrained models, strategies and datasets.

The DeBERTa-v3-large model outperforms about 5% the other two pretrained models we used in this task. Many other tasks show the outperformance of the DeBERTa model as well. DeBERTa modes proposed two novel tricks to improve the ability to solve many natural language tasks. Compared with the BERT and RoBERTa models, the disentangled attention mechanism is applied to show each word in two vectors, which represent its content and relative position. The disentangled matrices are used to calculate the attention weights between the word content and position. An enhanced mask decoder is adopted to help with model pre-training by using the absolute position to predict the masked tokens. The DeBERTa model also shows a big improvement in how well the model generalizes when the virtual adversarial training method is used.

Multi-sample dropout is a regularization technique which can accelerate training convergence and improve the model generalization compared to the network structure with a traditional dropout layer (Inoue, 2019). Four dropout layers were applied to the pooling layer output from the pretrained model. Table 2 shows that additional multi-sample dropout layer provides a remarkable effect (4% improvement on F1 score) on this classification task.

Moreover, many effective adversarial training and vitual adversarial training methods can improve the model robustness and the regularization on classification tasks (Madry et al., 2017; Miyato et al., 2016; Goodfellow et al., 2014; Zhu et al., 2019; Jiang et al., 2019; Qin et al., 2019; Shafahi et al., 2019). In this task, we adopted Projected Gradient Descent (PGD) and the Fast Gradient Method (FGM) to implement the perturbation on sequence embedding in this task (Madry et al., 2017; Miyato et al., 2016). Table 2 shows a pretty similar F1 score for both methods. The FGM adds small perturbations to the embedding layers to enhance the quality of word embedding. The submitted models are fused by models trained in the fast gradient method, credited with its fast training converges and fewer computation resources.

Additionally, the constrastive loss is considered in our training progress. It maximizes the amount of agreement between different augmented views of the same dataset through adding a contrastive loss in the latent space. (Hadsell et al., 2006; Chen et al., 2020; He et al., 2020a). Many tasks are competitively performed by adding contrastive loss functions, such as triple margin loss, NPair loss, InfoNCE loss, and SupCon loss (Sohn, 2016; Chen et al., 2020; Van den Oord et al., 2018; He et al., 2020a; Khosla et al., 2020). In this task, we considered the NTXent and SupCon losses as the additional contrastive loss. We added the selected contrastive loss to the cross-entropy loss to improve the classification accuracy.

The effect of the contrastive temperature reflects the attention of difficult samples. The smaller temperature pays more attention to the separation of the sample from the most similar one to it (Wang and Liu, 2021). We tuned the contrastive temperature in both NTXent and SupCon loss to train different classification models. NTXent with a contrastive temperature of 0.2 creates the best performance on the competition blind test set according to Table 2.

Furthermore, the weighted voting on multiple models with different pretrained models and random seeds improves final performance on the competition's blind test set. Figure 3 shows the minimum, mean, and maximum F1, precision and recall scores for predictions from 15 single models. The mixed voting method we applied for the final sub-

Dataset	Pre-trained model	Adversarial Training method	Text Features	Multi-sample dropout	Contrastive loss/temp	F1 @ Sarcasm	Macro F1
SemEval2022(EN)	roberta-large	pgd	False	False	NTXent/0.5	0.4125	0.6402
		pgd	False	True	NTXent/0.5	0.4582	0.6794
		fgm	False	True	NTXent/0.5	0.4691	0.6788
		fgm	False	True	SupCon/0.5	0.4788	0.6802
		fgm	False	True	SupCon/0.2	0.4813	0.6833
		fgm	True	True	NTXent/0.2	0.4862	0.6957
	deberta-v3-large	fgm	False	True	NTXent/0.2	0.5370	0.7132
SemEval2022 (EN+AR)	xlm-roberta-large	fgm	False	True	NTXent/0.2	0.3871	0.6420
SemEval2022(EN) + Semeval2018 + iSarcasm	roberta-large	fgm	False	True	NTXent/0.2	0.5570	0.7374
SemEval2022(EN) + Semeval2018 + iSarcasm + multi-model	deberta-v3-large	fgm	False	True	NTXent/0.2	0.5882	0.7445
	xlm-roberta-large	fgm	False	True	NTXent/0.2	0.5615	0.7409
SemEval2022 (EN+AR) + Semeval2018 + iSarcasm + multi-model	xlm-roberta-large	fgm	True	True	NTXent/0.2	0.6029	0.7676

Table 2: F1 scores using different strategies and datasets

mission has obtained the best performance.

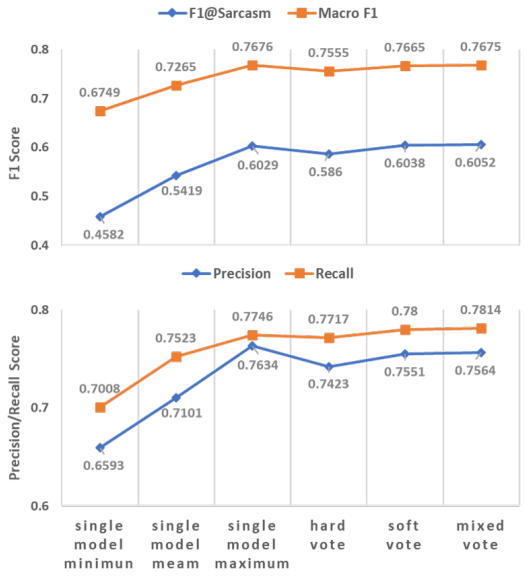

Figure 3: Prediction results from single and fused models

4 Conclusion

According to the performance on the blind test set, the proposed model with the highest F1 score in the sarcastic category applied four layers of multi-sample dropout with a rate of 0.4 following the pooling layer outputting from the XLM-RoBERTa-large model. The model is trained in the fast gradient method using the AdamW optimizer at a learning rate of 1e-05. The combination of the cross-entropy and the NTX contrastive loss is applied during the training process. Additionally, incorporating text features and data from other sources can help improve the prediction's accuracy.

Acknowledgements

This task has been completed with funding from PingAn Life Insurance. All the work stated in this paper was conducted during the Semeval-2022 competition. All the opinions and statements only reflect the authors' views and assertions. Addi-

tionally, we wish to express our gratitude to the task organizers and anonymous reviewers for their insightful comments.

References

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.

Santosh Kumar Bharti, Bakhtyar Vachha, RK Pradhan, Korra Sathya Babu, and Sanjay Kumar Jena. 2016. Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks*, 2(3):108–121.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 42–49.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Ruth Filik, Alexandra Ţurcan, Dominic Thompson, Nicole Harvey, Harriet Davies, and Amelia Turner. 2016. Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69(11):2130–2146.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020a. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020b. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Irazú Hernández-Farías, José-Miguel Benedí, and Paolo Rosso. 2015. Applying basic features from sentiment analysis for automatic irony detection. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 337–344. Springer.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Avinash Kumar, Vishnu Teja Narapareddy, Veerubhotla Aditya Srikanth, Aruna Malapati, and Lalita Bhanu Murthy Neti. 2020. Sarcasm detection using multi-head attention based bidirectional lstm. *Ieee Access*, 8:6388–6397.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

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*. Association for Computational Linguistics.

Reynier Ortega-Bueno, Francisco Rangel, D Hernández Farıas, Paolo Rosso, Manuel Montes-y Gómez, and José E Medina Pagola. 2019. Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org*, volume 2421, pages 229–256.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. *arXiv preprint arXiv:1610.08815*.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. 2019. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 32.

Antonio Reyes and Paolo Rosso. 2014. On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614.

Dymitr Ruta and Bogdan Gabrys. 2005. Classifier selection for majority voting. *Information fusion*, 6(1):63–81.

Samer Muthana Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, and Bianca Wright. 2020. Sarcasm detection using machine learning algorithms in twitter: A systematic review. *International Journal of Market Research*, 62(5):578–598.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Devendra Kr Tayal, Sumit Yadav, Komal Gupta, Bhawna Rajput, and Kiran Kumari. 2014. Polarity detection of sarcastic political tweets. In *2014 International conference on computing for sustainable global development (INDIACom)*, pages 625–628. IEEE.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.

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.

Palak Verma, Neha Shukla, and AP Shukla. 2021. Techniques of sarcasm detection: A review. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 968–972. IEEE.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Hamed Yaghoobian, Hamid R Arabnia, and Khaled Rasheed. 2021. Sarcasm detection: A comparative study. *arXiv preprint arXiv:2107.02276*.

Shiwei Zhang, Xiuzhen Zhang, Jeffrey Chan, and Paolo Rosso. 2019. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5):1633–1644.

Yong Zhang, Hongrui Zhang, Jing Cai, and Binbin Yang. 2014. A weighted voting classifier based on differential evolution. In *Abstract and Applied Analysis*, volume 2014. Hindawi.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.