TUG-CIC at SemEval-2021 Task 6: Two-stage Fine-tuning for Intended Sarcasm Detection

Jason Angel
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
ajason08@gmail.com

Segun Taofeek Aroyehun
TU Graz
Graz, Austria
aroyehun.segun@gmail.com

Alexander Gelbukh
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
www.gelbukh.com

Abstract

We present our systems and findings for the iSarcasmEval: Intended Sarcasm Detection In English and Arabic at SEMEVAL 2022. Specifically we take part in the Subtask A for the English language. The task aims to determine whether a text from social media (a tweet) is sarcastic or not. We model the problem using knowledge sources, a pre-trained language model on sentiment/emotion data and a dataset focused on intended sarcasm. Our submission ranked third place among 43 teams. In addition, we show a brief error analysis of our best model to investigate challenging examples for detecting sarcasm.

1 Introduction

According to the Freedictionary[1], sarcasm is a cutting statement to express contempt or ridicule, often using words to convey a meaning that is the opposite of their literal or actual meaning.

Due to its ambiguous nature, sarcasm plays an important role for resolution of several NLP tasks, such as Sentiment Analysis (Liu et al., 2010; Joshi et al., 2017; Maynard and Greenwood, 2014), Hate speech (Frenda et al., 2022), disagreement classification (Ghosh et al., 2021), Opinion Mining (Kannangara, 2018) among others. One example of benefits of modeling sarcasm is presented by Bouazizi and Ohtsuki (2015) about the use of sarcastic tweets to improve Sentiment Analysis.

In this work we present our submission to iSarcasmEval at Semeval-2022 for subtask A (Given a text, determine whether it is sarcastic or non-sarcastic). Our solution system is at the top three teams among 43 teams. The proposed approach uses the pre-trained language model well informed of this task, because it uses sentiment/emotion data and we enhanced it with a dataset about intended sarcasm texts.

The rest of the document goes as follows: Section 2 overviews some related works about sarcasm detection tasks. In Section 3, we give a detailed description of our conducted experiments, including the dataset used. Section 4 summarizes our results and offers an interpretation to our findings. Finally, Section 5 presents our conclusion, contributions and future work.

2 Related works

Beside the importance of detecting sarcasm for industry applications related to understanding comments on social networks or commercial products reviews (Yavanoglu et al., 2018), sarcasm detection has received great attention from the NLP community, influencing the creation of diverse approaches, benchmark datasets and evaluation methods.

The methods for modeling sarcasm range from rule-based system and statistical approaches by exploiting handcrafted features using traditional machine learning (Joshi et al., 2015; Wicana et al., 2017) to modern techniques using deep learning architectures together with word embeddings and pretrained language models (Mehndiratta and Soni, 2019; Srivastava et al., 2020; Wang et al., 2021).

Previous attempts for benchmarking sarcasm detection were proposed by "SemEval-2018 Task 3: Irony Detection"[2], ALTA Shared Task 2019[3], FigLang 2020[4] and WANLP 2021[5]. They crafted their datasets either using weak supervision tech-

[1] https://www.thefreedictionary.com/

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

[3] http://www.alta.asn.au/events/sharedtask2019/description.html

[4] https://sites.google.com/view/figlang2020/shared-tasks

[5] https://sites.google.com/view/wanlp2021

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 951 - 955

niques like scraping tweets having the #irony, #sarcasm hashtags, or by manual labeling from third party annotators, which leads to several shortcomings as exposed in Oprea and Magdy (2020), instead the dataset used in this work (iSarcasm) was crafted by asking the authors themselves to provide the sarcastic/non-sarcastic labeling for their tweets.

	Training	validation
Sarcastic	867	86
Non-Sarcastic	2601	259
Total examples	3468	345

Table 1: Number of classes in train and validation splits

3 Experiments

Our experiments aim to determine whether a given text is intended to be sarcastic or non-sarcastic. The metric for ranking systems on the competition is the F1-score for the sarcastic class (positive label) only.

Datasets. As part of the iSarcasm shared task (Abu Farha et al., 2022), the organizers provide training and test datasets for English and Arabic, however, we only participated in the English track. Table 1 displays the number of samples and the distribution over sarcastic and non-sarcastic texts for train and validation splits. As observed sarcastic texts roughly account for the 25% of data in each split.

In addition to the iSarcasmEval dataset, we also employed the SPIRS dataset[6] which is a collection of sarcastic and non-sarcastic tweet ids (15,000 for each category) gathered using "reactive supervision", a new data capturing method (Shmueli et al., 2020). From SPIRS only intended and negative examples were considered which amount to 15,950 and 1,773 examples in the training and development splits, respectively. Therefore, tweets perceived as sarcastic were discarded, since perceived sarcasm accounts for a related but different task.

Pre-processing steps. We perform minimal preprocessing which includes conversion of user mentions and links to @USER and URL, respectively. Also, consecutive whitespaces are normalized to a single occurrence and punctuation marks are surrounded with a single space character.

Our approach. Our approach is informed by insights from the literature on the effectiveness of sentiment/emotion information for sarcasm detection. Additionally, the creators of the SPIRS dataset also provided perspectives (intended vs. perceived) for their heuristically-labeled sarcasm dataset. Building on these, we examine the effect of these two knowledge sources on the sarcasm detection task. Specifically, we combine pre-training on sentiment/emotion with a second pre-training step on the intended sarcasm subset of the SPIRS dataset. In the final step, we fine-tune the pre-trained model on the dataset provided for the shared task.

Pre-trained models. We employed four publicly available pre-trained models from the Huggingface model hub. The models namely: BERTweet-base[7], BERTweet-sentiment[8], BERTweet-emotion[9], and BERTweet-large[10] are trained on twitter data using masked language modeling as well as sentiment and emotion detection where applicable.

Training details/hyperparameters. We use the adapter-transformers library for the experiments. We add a classification layer (consisting of two dense layers) on top of the pooled output of the last transformer layer and optimize this layer jointly with the pre-trained layers. We optimize the model using Adamw with a batch size of 64 on a single Nvidia V100 GPU (32GB) and a maximum learning rate of 1e-5. We use a warmup ratio of 0.1 and set the maximum number of epochs to 15 with earlystopping on the validation performance metric (F1) using a patience of 5 evaluation runs. We evaluate the performance of the model every 20 steps on the validation set. Furthermore, we employ three regularization techniques: weight decay with a factor of 0.01, dropout applied to the pooled output of the last transformer layer with a probability of 0.2, and label smoothing with a factor of 0.1.

[6] https://github.com/bshmueli/SPIRS

[7] https://huggingface.co/vinai/bertweet-base

[8] https://huggingface.co/cardiffnlp/bertweet-base-sentiment

[9] https://huggingface.co/cardiffnlp/bertweet-base-emotion

[10] https://huggingface.co/vinai/bertweet-large

Model	Validation			Test		
	P	R	F1	P	R	F1
BERTweet-emotion-spirs	0.515	0.581	0.546	0.323	0.710	0.444
BERTweet-sentiment-spirs	0.545	0.558	0.552	0.330	0.705	0.450
BERTweet-spirs	0.526	**0.593**	0.557	0.349	0.710	0.468
BERTweet-large-spirs	**0.667**	0.558	**0.608**	**0.412**	**0.740**	**0.530**
Post-evaluation runs*						
BERTweet	0.573	0.547	0.560	0.395	0.680	0.500
BERTweet-emotion	0.529	0.535	0.532	0.378	0.630	0.473
BERTweet-sentiment	0.527	0.570	0.547	0.348	0.675	0.459
BERTweet-large	**0.690**	0.570	**0.624**	**0.420**	0.735	**0.535**
BERTweet-irony	0.581	0.581	0.581	0.399	0.665	0.499
BERTweet-irony-spirs	0.515	0.581	0.546	0.323	0.710	0.444

Table 2: Performance scores on the validation and test sets (All metrics are for the sarcastic class).
*Suggested by one of the reviewers for completeness and better comparison.

Case	Example	Explanation
Not enough context to determine the intention of the phrase(s)	- JUSTICE HAS BEEN DONE . - i ' ve never had protected sex - i'm dying	Chances are that the required context are on other parts of the post, e.g. in the comments
Common sense or Knowledge of real world is required	Mad how many cars they make now without indicators [happy-emoji]	It is not that new cars had a design without indicators, it is just that for many reasons, some people refuse to use them, and that make other drivers angry because of the difficulties while driving.
Common sense or Knowledge of real world is required	Vaccinated this morning . Not sure it ' s worked - still committed to Apple and no extra autism detected . Bloody science .	Those 'secondary effects' are not even related with vaccines; it seems that the user is just joking with that concept.
Common sense or Knowledge of real world is required	hello to all three of my followers, this is my big return to twitter	People who are familiar with Twitter knows that having three followers is not actually impressive. Instead this author is making fun of it.
Sentiment contradiction between key words and emojis	- I hate it here [happy emoji] - Headaches that last all day nonstop [happy emoji] - @ user thanks for shutting down the only Disney Store in northeastern Ohio . The other two stores went out of business yrs ago . It ' s not the same with shipping costs online not to mention LE doll sales are an absolute nightmare online . So unbelievably disappointing . [sad emoji]	Not all examples use positive emoji to end sarcastic messages. Here, the contradiction lies in the pair thanks-[sad emoji]

Table 3: Error analysis of model predictions on the validation set

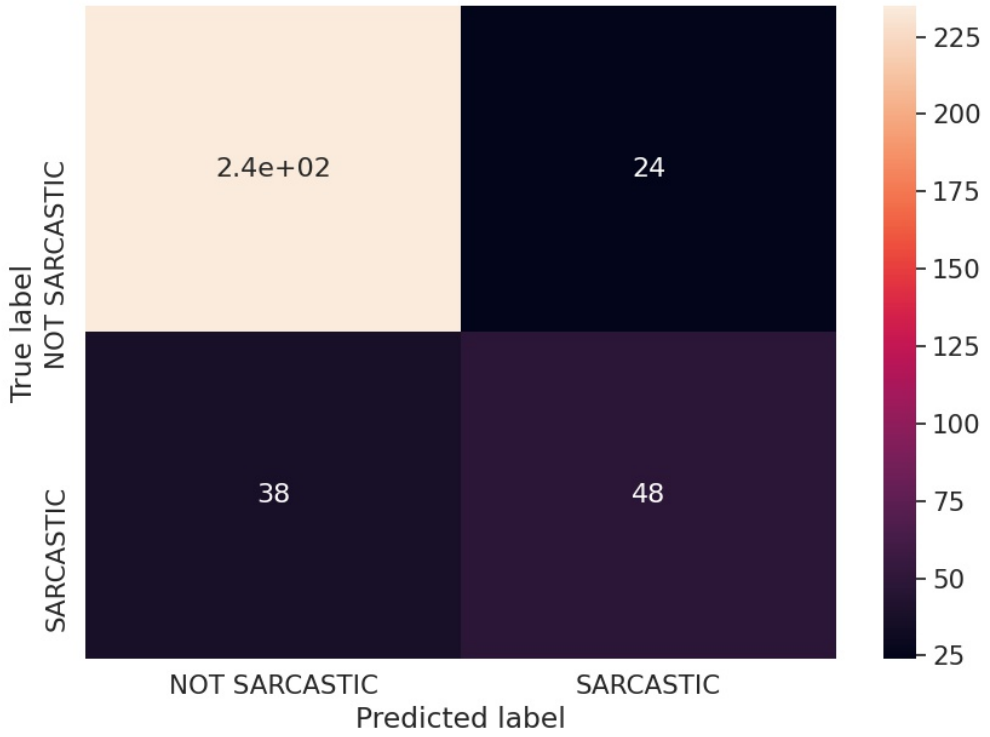

Figure 1: Confusion matrix for the best model predictions on the validation set

4 Results and error analysis

Our system ranks third on the leaderboard for the competition with an F1-score for the sarcastic class (positive label) of 0.530. The performance scores for our submissions on the validation and test sets are in Table 2. The BERTweet-large model achieved the best score out of our four submissions. It seems that pre-training on sentiment or emotion detection is not beneficial in our experimental settings. This observation can be confirmed in our post-evaluation runs where the BERTweet-base model shows superior performance (up to 5 F1 points on the test set) when compared to the same model that has been further pre-trained on sentiment or emotion detection task. Further investigation of this observation is required to identify alternative approaches to incorporate these sources of information for transformer-based models. Figure 1 shows the confusion matrix of the predictions made by our best model on the validation set.

As for error analysis, we use the validation set to identify possible explanations for why the model fails at predicting correctly positive labels (sarcastic tweets). Table 3 shows three cases as challenges for detecting sarcasm, which are usually hard to overcome in many NLP tasks as well.

These results show that still modern language model techniques struggle to grasp the pragmatics of messages expressed in social media, which are particularly difficult.

5 Conclusion

We present our systems for the task of sarcasm detection using a variety of knowledge sources that explore the capabilities of pre-trained language models to understand pragmatic phenomena such as sarcasm. We found that although these knowledge sources help to shed light over the sarcasm detection task, still more external knowledge would be required to correctly classify difficult cases that require a deeper understanding of real world context.

Based on our analysis, we plan to further examine the impact of including say emoji modelling to measure its influence, especially over cases that show a contradiction on expressed sentiments.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico, grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

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.

Mondher Bouazizi and Tomoaki Ohtsuki. 2015. Opinion mining in twitter how to make use of sarcasm to enhance sentiment analysis. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1594–1597.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, page 116398.

Debanjan Ghosh, Ritvik Shrivastava, and Smaranda Muresan. 2021. " laughing at you or with you": The role of sarcasm in shaping the disagreement space. *arXiv preprint arXiv:2101.10952*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. 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.

Sandeepa Kannangara. 2018. Mining twitter for fine-grained political opinion polarity classification, ideology detection and sarcasm detection. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 751–752.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Diana G Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Lrec 2014 proceedings*. ELRA.

Pulkit Mehndiratta and Devpriya Soni. 2019. Identification of sarcasm using word embeddings and hyperparameters tuning. *Journal of Discrete Mathematical Sciences and Cryptography*, 22(4):465–489.

Silviu Vlad Oprea and Walid Magdy. 2020. The effect of sociocultural variables on sarcasm communication online. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–22.

Boaz Shmueli, Lun-Wei Ku, and Soumya Ray. 2020. Reactive Supervision: A New Method for Collecting Sarcasm Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2553–2559, Online. Association for Computational Linguistics.

Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava. 2020. A novel hierarchical bert architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97.

Haiyang Wang, Xin Song, Bin Zhou, Ye Wang, Liqun Gao, and Yan Jia. 2021. Performance evaluation of pre-trained models in sarcasm detection task. In *International Conference on Web Information Systems Engineering*, pages 67–75. Springer.

Setra Genyang Wicana, Taha Yasin İbisoglu, and Uraz Yavanoglu. 2017. A review on sarcasm detection from machine-learning perspective. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 469–476. IEEE.

Uraz Yavanoglu, Taha Yasin Ibisoglu, and Setra Genyang Wıcana. 2018. Sarcasm detection algorithms. *International Journal of Semantic Computing*, 12(03):457–478.