LISACTeam at SemEval-2022 Task 6: A Transformer based Approach for Intended Sarcasm Detection in English Tweets

Abdessamad Benlahbib[1]*, **Hamza Alami**[2]*, **Ahmed Alami**[3]
[1] LISAC Laboratory, Faculty of Sciences Dhar EL Mehraz, USMBA, Fez, Morocco
[2] Innov-Tech Laboratory, Department of Engineering, High Technology School, Rabat
[3] Laboratory of Engineering Sciences, National School of Applied Sciences,
Ibn Tofail University, Kenitra, Morocco
abdessamad.benlahbib@usmba.ac.ma , hamza0alami@gmail.com
alami.alami1996@gmail.com

Abstract

In this paper, we present our system and findings for SemEval-2022 Task 6 - iSarcasmEval: Intended Sarcasm Detection in English. The main objective of this task was to identify sarcastic tweets. This task was challenging mainly due to (1) the small training dataset that contains only 3468 tweets and (2) the imbalanced class distribution (25% sarcastic and 75% non-sarcastic). Our submitted model (ranked eighth on Sub-Task A and fifth on Sub-Task C) consists of a Transformer-based approach (BERTweet model).

1 Introduction

The Cambridge Dictionary [1] defines sarcasm as *"the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way."*

Due to the Web openness, sarcastic content becomes very frequent in social media and e-commerce platforms, which may cause misunderstandings. Furthermore, identifying such content is a very challenging task even for humans (Farias and Rosso, 2017). Also, it could impact some natural language processing tasks such as sentiment analysis (Farias and Rosso, 2017; Do et al., 2019; Tubishat et al., 2018; Balazs and Velásquez, 2016; Maynard and Greenwood, 2014; Ptáček et al., 2014; Bouazizi and Otsuki Ohtsuki, 2016; Ren et al., 2018).

We introduce the following example: *"The movie was enjoyable to the point that I clapped because it is finished."* For an opinion mining system, this sentence could be considered positive. However, the author expresses a negative judgment against the movie since the expression *"I clapped because it is finished"* means that it was boring.

For this reason, SemEval 2022 set up Task 6: iSarcasmEval - Intended Sarcasm Detection in English and Arabic to detect sarcastic and non-sarcastic tweets (Abu Farha et al., 2022). Our submitted system consisted of a pre-trained transformer model for English Tweets named BERTweet (Nguyen et al., 2020), secured 8th and 5th positions respectively on Sub-Task A and Sub-Task C leaderboard.

The rest of the paper is structured in the following manner: Section 2 provides the data structure and the main objective of each Sub-Task. Section 3 describes our system. Section 4 details the experiments. And finally, Section 5 concludes this paper.

2 Task Description

The organizers of this task introduced two tweet datasets for both English and Arabic languages that contain:

- a label specifying whether a tweet is sarcastic or non-sarcastic, provided by its author.
- a non-sarcastic rephrase of a sarcastic tweet provided by its author.
- a label specifying the category of ironic speech that it reflects, provided by a linguistic expert (English only).
- a label specifying the dialect (Arabic only).

This task consists mainly of three sub-tasks for the English dataset and two sub-tasks for the Arabic dataset where Sub-Task A aims at determining whether a tweet is sarcastic or non-sarcastic, Sub-Task B, which is available for English only, is a binary multi-label classification task that intends to determine which ironic speech category a sarcastic tweet belongs to if any, and finally, Sub-Task C that takes two inputs: a sarcastic tweet and its non-sarcastic rephrase, and focuses on identifying the sarcastic one between them.

*contributed equally
[1] https://dictionary.cambridge.org/fr/dictionnaire/anglais/sarcasm

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 993 - 998

3 System Description

In this section, we describe our proposed system that tackles Sub-Task A and Sub-Task C English.

3.1 Sub-Task A

In order to tackle Sub-Task A, we adopted a transformer-based (Vaswani et al., 2017) approach that consists of fine-tuning BERTweet [2], which is a language model pre-trained on 850M English Tweets, and it has the same architecture as BERT-base (Devlin et al., 2019), as well as it is was trained using the RoBERTa pre-training procedure (Liu et al., 2019).

Before feeding the training data to BERTweet model, we preprocessed them by removing URLs and then replacing emojis with their English textual meaning (Alami et al., 2020) using BERTweet demojizer [3]. Figure 1 depicts the Tweets preprocessing pipeline.

After the preprocessing phase, we fine-tuned BERTweet model on the training dataset that contains 3468 tweets (867 sarcastic tweets and 2601 non-sarcastic tweets).

3.2 Sub-Task C

The same model of Sub-Task A was used to handle Sub-Task C by feeding two texts to the BERTWeet model that was already fine-tuned on the training dataset. The text with the highest probability of being sarcastic is considered the sarcastic one.

4 Experimental Results

We experimented our model on the SemEval 2022 Task 6: iSarcasmEval - Intended Sarcasm Detection in English Sub-Task A and Sub-Task C datasets. All our experiments have been conducted in Google Colab environment[4], The following libraries: Transformers - Hugging Face[5] (Wolf et al., 2020), Scikit-Learn[6] (Pedregosa et al., 2011), and Keras[7] were used to train and to asses the performance of our model.

[2] https://github.com/VinAIResearch/BERTweet

[3] https://huggingface.co/transformers/v4.4.2/_modules/transformers/models/bertweet/tokenization_bertweet.html

[4] https://colab.research.google.com/

[5] https://huggingface.co/docs/transformers/index

[6] https://scikit-learn.org/stable/

[7] https://keras.io/

4.1 Datasets

Since we have participated in Sub-Task A and Sub-Task C for English, we will only describe the English dataset. The training set contains 867 sarcastic tweets and 2601 non-sarcastic tweets, the test set of Sub-Task A contains 200 sarcastic tweets and 1200 non-sarcastic tweets, and the test set of Sub-Task C contains 200 sarcastic tweets and their rephrases. Figure 2 depicts the class distribution of the English tweets in the training and test set for Sub-Task A.

4.2 Evaluation Metric

To evaluate the performance of the submitted results, the organizers adopted the F1-score for the sarcastic class as the main metric for Sub-Task A as well as the accuracy for Sub-Task C. The F1-score and accuracy are computed in the following manner where $P_{sarcastic}$ and $R_{sarcastic}$ are respectively the precision and recall of the sarcastic class, and TP, TN, FP and FN are respectively the true positive, true negative, false positive and false negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$P_{sarcastic} = \frac{TP_{sarcastic}}{TP_{sarcastic} + FP_{sarcastic}} \quad (2)$$

$$R_{sarcastic} = \frac{TP_{sarcastic}}{TP_{sarcastic} + FN_{sarcastic}} \quad (3)$$

$$F1_{sarcastic} = \frac{2 \times P_{sarcastic} \times R_{sarcastic}}{P_{sarcastic} + R_{sarcastic}} \quad (4)$$

4.3 Experimental Settings

During the fine-tuning of BERTWeet model, we set the hyper-parameters as follows: 10^{-5} as the learning rate, 15 epochs, 128 as the max sequence length, and 32 as batch size. The same settings were adopted for DistilBERT (Sanh et al., 2019) and BERT base uncased. Table 1 summarizes the hyperparameters settings of BERTWeet model.

For the Bidirectional Long Short-Term Memory (Bi-LSTM) (Hochreiter and Schmidhuber, 1997) and Bidirectional Gated Recurrent Unit (Bi-GRU) (Cho et al., 2014), we set 10 epochs, 128 as the max sequence length, and 16 as batch size.

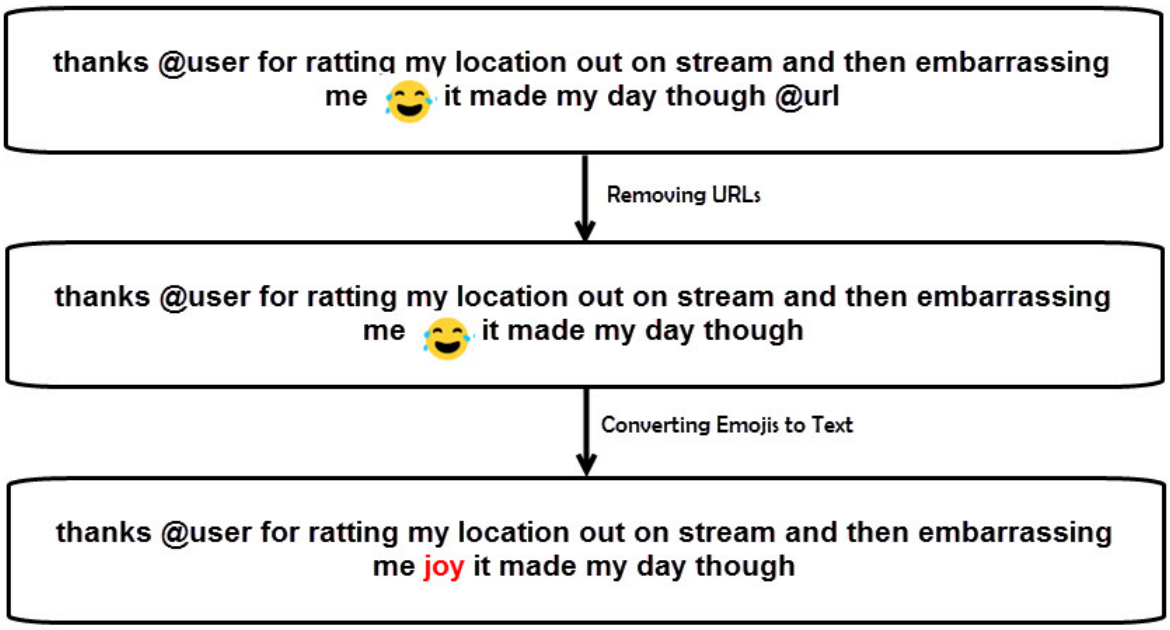

Figure 1: Tweets preprocessing

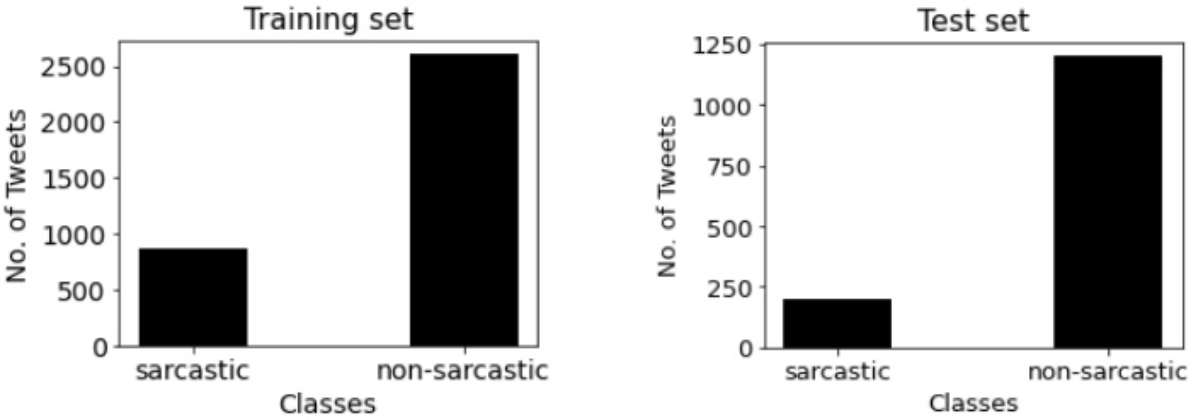

Figure 2: Class distribution of the English tweets in the training and test set for Sub-Task A

Hyperparameters	Settings
Learning rate	10^{-5}
Batch size	32
Epochs	15
Max sequence length	128
Optimizer	Adam (Kingma and Ba, 2015)
Loss	Cross-Entropy

Table 1: Hyperparameters settings for BERTWeet model in the experiments

4.4 System Performance

We evaluated various models on Sub-Task A test set including Linear Support Vector Classification (LinearSVC) (Boser et al., 1992), Logistic Regression, Multinomial Naive Bayes (MultinomialNB), Bi-LSTM, Bi-GRU, DistilBERT, BERT base uncased, RoBERTa base, and BERTweet base. We picked the combination of unigrams, bigrams, and trigrams of token counts as features for LinearSVC, Logistic Regression, and MultinomialNB since this combination delivered the best results in terms of the F-1 sarcastic metric.

For non-transformer-based models, we preprocessed the data by removing stop words and special characters. For transformer-based models, two approaches were adopted during the evaluation phase. In the first approach, we preprocessed the data as described in Figure 1. In the second one, we fine-

tuned the model without applying any preprocessing to the data. Table 2 depicts the obtained results of various models on Sub-Task A - English.

We can see from Table 2 that BERTweet base model achieved the best results in detecting sarcastic tweets succeeded by RoBERTa base. Surprisingly, LinearSVC achieved better results than BERT base and DistilBERT.

We evaluated various models on Sub-Task C test set including LinearSVC, Logistic Regression, MultinomialNB, Bi-LSTM, Bi-GRU, RoBERTa base, and BERTweet base. Table 3 depicts the obtained results of various models on Sub-Task C - English. We mention that the same preprocessing approaches applied on Sub-Task A tweets were applied on Sub-Task C test set.

According to the reported results in Table 3, we can see that BERTweet base model achieved the best results succeeded by RoBERTa base. Moreover, we notice that traditional machine learning approaches such as LinearSVC, Logistic Regression, and MultinomialNB outperformed Recurrent Neural Networks: Bi-LSTM and Bi-GRU.

5 Conclusion

In this paper, we described our approach for tackling Sub-Task A and Sub-Task C of SemEval 2022 Task 6: iSarcasmEval - Intended Sarcasm Detection in English. Our submitted system consisted of a pre-trained transformer model for English Tweets named BERTweet, secured 8th and 5th positions respectively on Sub-Task A and Sub-Task C leaderboard.

Since the top-ranked system for the English Sub-Task A scored about 0.6052 F1-score for the sarcastic class, future studies and works will focus on improving the performance of sarcasm detection tasks by adopting other approaches such as data augmentation and oversampling.

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.

Hamza Alami, Said Ouatik El Alaoui, Abdessamad Benlahbib, and Noureddine En-nahnahi. 2020. LISAC FSDM-USMBA team at SemEval-2020 task 12: Overcoming AraBERT's pretrain-finetune discrepancy for Arabic offensive language identification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2080–2085, Barcelona (online). International Committee for Computational Linguistics.

Jorge A. Balazs and Juan D. Velásquez. 2016. Opinion mining and information fusion: A survey. *Information Fusion*, 27:95–110.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA. Association for Computing Machinery.

Mondher Bouazizi and Tomoaki Otsuki Ohtsuki. 2016. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. 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.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118:272–299.

D.I. Hernández Farias and P. Rosso. 2017. Chapter 7 - irony, sarcasm, and sentiment analysis. In Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu, editors, *Sentiment Analysis in Social Networks*, pages 113–128. Morgan Kaufmann, Boston.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Sub-Task A - English					
Model	**F1-sarcastic**	F1-score	Precision	Recall	Accuracy
LinearSVC	0.3508	0.6055	0.6304	0.5957	0.7700
Logistic Regression	0.3076	0.6021	0.5950	0.6124	0.8200
MultinomialNB	0.2982	0.5924	0.5904	0.5948	0.8050
Bi-LSTM	0.2667	0.5575	0.5700	0.5536	0.7486
Bi-GRU	0.2358	0.5330	0.5463	0.5329	0.7221
DistilBERT (with preprocessing)	0.3070	0.5861	0.5975	0.5799	0.7743
DistilBERT (without preprocessing)	0.3267	0.5809	0.6162	0.5758	0.7350
BERT base (with preprocessing)	0.3388	0.5888	0.6258	0.5823	0.7407
BERT base (without preprocessing)	0.3087	0.5887	0.5983	0.5829	0.7793
RoBERTa base (with preprocessing)	0.3746	0.6454	0.6262	0.6823	0.8521
RoBERTa base (without preprocessing)	0.3984	0.6351	0.6642	0.6218	0.7886
BERTweet base (with preprocessing) (Official Submission)	0.4291	0.6513	0.6353	0.6896	0.7929
BERTweet base (without preprocessing)	**0.4334**	0.6547	0.6917	0.6384	0.7964

Table 2: The obtained results of various models on Sub-Task A - English test set

Sub-Task C - English				
Model	F1-score	Precision	Recall	**Accuracy**
LinearSVC	0.5392	0.5657	0.5952	0.5850
Logistic Regression	0.5996	0.6008	0.6004	0.6000
MultinomialNB	0.6033	0.6034	0.6033	0.6050
Bi-LSTM	0.4834	0.4835	0.4836	0.4850
Bi-GRU	0.5200	0.5233	0.5233	0.5200
RoBERTa base (with preprocessing)	0.7177	0.7172	0.7188	0.7200
RoBERTa base (without preprocessing)	0.7186	0.7186	0.7186	0.7200
BERTweet base (with preprocessing) (Official Submission)	0.7737	0.7735	0.7740	**0.7750**
BERTweet base (without preprocessing)	0.7585	0.7581	0.7589	0.7600

Table 3: The obtained results of various models on Sub-Task C - English test set

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).

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yafeng Ren, Donghong Ji, and Han Ren. 2018. Context-augmented convolutional neural networks for twitter sarcasm detection. *Neurocomputing*, 308:1–7.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Mohammad Tubishat, Norisma Idris, and Mohammad A.M. Abushariah. 2018. Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing & Management*, 54(4):545–563.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.