UoR-NCL at SemEval-2022 Task 6: Using ensemble loss with BERT for intended sarcasm detection

Emmanuel Osei-Brefo
University of Reading
White Knights Campus
Reading-UK
e.osei-brefo@student.reading.ac.uk

Huizhi Liang
University of Newcastle
Newcastle
UK
huizhi.liang@newcastle.ac.uk

Abstract

Sarcasm has gained notoriety for being difficult to detect by machine learning systems due to its figurative nature. In this paper, Bidirectional Encoder Representations from Transformers (BERT) model has been used with ensemble loss made of cross-entropy loss and negative log-likelihood loss to classify whether a given sentence is in English and Arabic tweets are sarcastic or not. From the results obtained in the experiments, our proposed BERT with ensemble loss achieved superior performance when applied to English and Arabic test datasets. For the validation dataset, our model performed better on the Arabic dataset but failed to outperform the baseline method (made of BERT with only a single loss function) when applied on the English validation set.

1 Introduction

In recent times, Social media platforms such as Twitter have turned out to be one of the most influential media for the expression of views, emotions, and information. To characterize sarcasm on Twitter; (Parmar et al., 2018) proposed the following: (a) strife between negative situation and positive sentiment, (b) strife between positive situation and negative sentiment, (c) Tweet starts with an interjection word, (d) likes and dislikes contradiction, (e) tweet conflicting ubiquitous facts, (f) tweets with positive sentiment and antonym pair, and (g) tweet contrasting facts that are time-sensitive.

With a huge amount of content being churned on social media and the need to analyze and detect sarcasm closely, text classification methods have been widely introduced to deal with these sophisticated tasks. Sarcasm has been shown to pose a major difficulty for sentiment analysis models (Liu, 2010), mainly because sarcasm acts as a form of verbal irony which enables the concealment of the true intention of denigration and negativity under a pretence of open a respectful representation such as; “The only thing I got from college is a caffeine addiction". As a result, the ability to detect sarcasm is crucial for understanding the real intents and beliefs of people (Maynard and Greenwood, 2014).

To address these issues and limitations, (Abu Farha et al., 2022) have organized the iSarcasmEval shared task for intended Sarcasm Detection In English and Arabic where different tweets in English and Arabic languages have to be classified as either sarcastic or not.

Several models for detection of sarcasm have been proposed in literature which incorporate statistical, machine learning, and rule-based approaches but predominantly (Mandal and Mahto, 2019). However, these techniques are not able to effectively perceive the figurative and ironic meaning of words (Joshi et al., 2016). Also, these methods require manually engineered features and are unable to understand the patterns in passive voice sentences (Bajwa and Choudhary, 2006). However, instead of utilizing manually engineered features, the incorporation of transformer models has the ability to automatically learn the important features.

In this paper, we present our participating system to the iSarcasmEval shared task (Abu Farha et al., 2022). In this work, we seek to apply a novel transformer-based approach to detect sarcastic statements in both English and Arabic languages and propose a novel approach to fine-tune the BERT model for sarcasm detection of both Arabic and En-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 871 - 876

glish tweets using ensemble loss. BERT is a model that is gaining increasing popularity due to its outstanding performance for multiple NLP tasks.

Recently, the language model called BERT has been widely used in several tasks. The pre-trained model can generate word/token or sentence representations that are enriched with prior knowledge. Then, they can be fine-tuned specifically for many downstream tasks such as sarcasm detection. The traditional BERT model only uses one loss function whilst our proposed model which combines a cross-entropy loss with a negative log-likelihood function leads to a better penalization of incorrect predictions that the model is confident about more than incorrect predictions the model is less confident about. This leads to better detection of sarcasm due to an improved F-1 sarcastic.

One of the main contributions of this work is the use of ensemble loss with BERT (Devlin et al., 2019) to address the task of sarcasm detection of tweets in both English and Arabic languages with improved sarcasm detection ability.

The rest of the paper is organized as follows: The Related Work in section 2 discusses the research works which are closely related to the current study. Section 3 contains the system description part and highlights the description of the proposed approach and its working methodology. The dataset used for experiment setup is provided in the Experiments section in section 4, whilst the results and Discussion segment can be found under section 4.1 and contains the experimental results and the performance metrics. The conclusion and future work for this research work are all captured in section 5.

2 Related Work

The conversation aspect in sarcasm detection was investigated by (Ghosh et al., 2017), and sarcastic and non-sarcastic tweets were collected to create a dataset for their experiments. BERT, which is a transformer-based machine learning model has been exploited to provide a deep contextual representation of words and used for sarcasm detection. They are pre-trained language models and have been also been widely utilized in many NLP domains. These models have been proven to be effective since they are enriched with knowledge from the pre-training resources (Avvaru et al., 2020). (Moores and Mago, 2022) recently surveyed automated sarcastic detection on Twitter and found out that there was a shift towards the use of deep learning methods due to the benefit of using a model with induced instead of discrete features combined with the innovation of transformers. BERT and GloVe embeddings were combined with machine learning classifiers to detect sarcasm in tweets in the work of (Khatri and P, 2020). Semi-supervised techniques were used by (Tsur et al., 2010) to detect sarcasm Sentences in Online Product Reviews whilst (Amir et al., 2016) also deployed the use of automatic learning and exploiting word embeddings to recognize sarcasm. Other approaches such as Bi-Directional Gated Recurrent Neural Network (Bi-Directional GRNU) have also been used to detect sarcasm (Zhang et al., 2016). Bert has been used with a crowd layer on tweets with noisy labels to achieve improved results (Osei-Brefo et al., 2021).

3 System description

For each sample in the English and Arabic dataset, a tweeted text and a given token were first concatenated in the following format:

[CLS] + Tweeted text + [SEP] sarcastic sentence [SEP]

where '[CLS]' token was added for classification and two '[SEP]' tokens were used to identify the sarcastic nature of the tweet. Each sentence was initially organized, after which the '[CLS]' and '[SEP]' tokens were added to their start and end. The tokenized sentence was then truncated or padded to a maximum length of 64. Each generated token was then mapped to its individual IDs to create an attention mask for each sentence. A hidden size of 768 was utilized for the BERT, whilst 120 and 2 were respectively used as the hidden size of the classifier and the number of labels. A one-layer

feed-forward classifier was then instantiated after which the last hidden state of the token '[CLS]', was extracted for classifying whether a statement was sarcastic or not.

Binary Classification for detecting sarcasm To fine-tune the models for sub-task A, a fully-connected layer was added on top of the pooled output (the sequence embedding from the pre-trained model). This layer had an output size of 2 with a softmax activation function which outputs the probability of each class as either 1 or 0. For sub-task A, the baseline model used was a Bert model with the same parameters as our proposed model except that it was made up of a single cross-entropy loss function. A cross-entropy loss and Negative log-likelihood loss functions were combined in different weights proportion to form the ensemble loss used as a loss function for fine-tuning. These proportional weight functions were 0.8/0.2 and 0.2/0.8 for the cross-entropy loss and Negative log-likelihood loss functions respectively as can be seen in Tables 1 and 2 as **BERT Ensemble loss A** and **BERT Ensemble loss B** respectively. The final prediction of each sentence was made by selecting the class with the maximum probability. Due to the difficulties inherent nature of detecting sarcasm by machines, a higher threshold of about 70 % was used as the minimum probability beyond which a tweet can be classified as being sarcastic for both the English and Arabic datasets.

Ensemble loss function Two loss functions were combined together in such a way that the benefits derived from each loss function were embedded in the model. These loss functions were cross-entropy loss and negative log-likelihood loss. The weights of these combinations were 0.8 to 0.2 and 0.2 to 0.8 for BERT Ensemble loss A and BERT Ensemble loss B respectively. On the other hand, the Baseline model loss was made up of only the cross-entropy loss.

4 Experiments

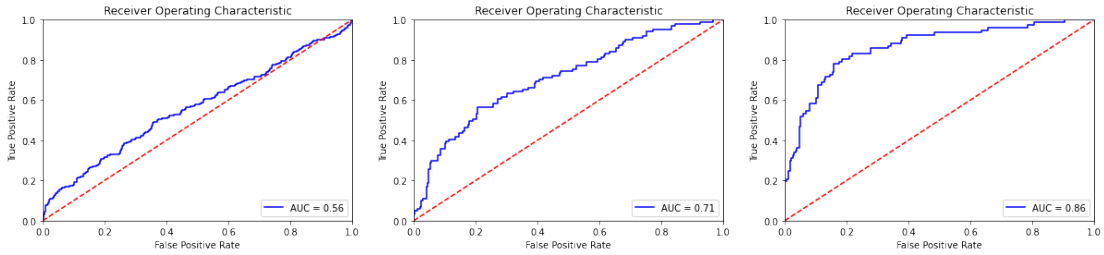
Experiments were conducted on the respective English and Arabic training and test sets provided by the task organizers. The training set of the English language had 3467 samples whilst that of the Arabic language had 3120 data samples. During the development stage, the training datasets were divided into 90% and 10% splits sets respectively for the training and validation datasets. The test sets for both languages on the other hand had a total of 1,400 data samples used in the evaluation phase.

All the models were fine-tuned by using the Adam optimizer for 12 epochs with the batch size 32 and the learning rate of 0.000005 and 0.000000001 as the default epsilon value. For sub-task A, the models were evaluated by Accuracy, Recall, false positives, F-1 sarcastic, Accuracy, and F1-macro.

4.1 Results and Discussion

Table 1 shows the results of the baselines and our approaches in both sub-task A using the Validation dataset. Table 2 also shows the results obtained using the test data sets

Our proposed model with the ensemble loss had different performance strengths on the validation dataset and the test datasets. Their relative performance also varied depending on the particular Language they were applied to. Among the 5 metrics used, the most effective metric that can be used to determine how accurate a model is able to detect sarcastic statements is the F-1 sarcastic metric. For the validation dataset, the baseline model outperformed our proposed model in all the 5 metrics used when applied to the English Language dataset. For the Arabic dataset, our model outperformed the baseline in all the other 3 metrics used with the exception of the false positive and f1-macro metrics where our model could not outperform the baseline. It is worth noting that our model had an F-1 sarcastic metric of 83.53% when it was applied to the Arabic language test dataset. For the test dataset, our proposed model outperformed the baseline model for both the English languages with an

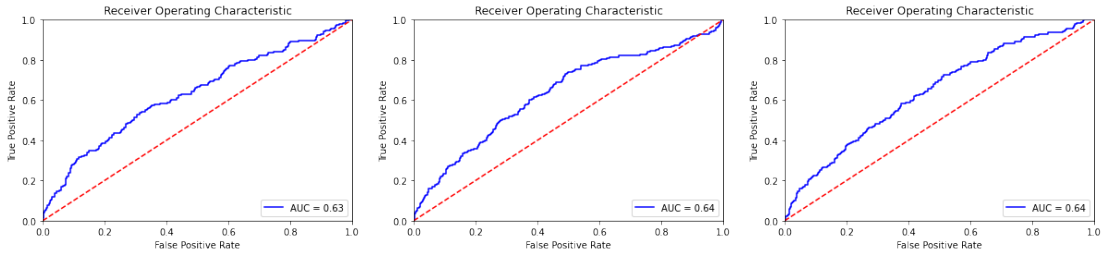


(a) Roc curve illustrating the performance of the baseline Bert model on test dataset English tweets

(b) Roc curve illustrating the performance of the Bert model on validation dataset English tweets

(c) Roc curve illustrating the performance of the Bert model on Test dataset Arabic tweets

Figure 1: Comparison of the proposed approach performance on on sub task A



(a) Roc curve illustrating the performance of the Bert Ensemble loss model on test dataset English tweets

(b) Roc curve illustrating the performance of the baseline Bert model on test dataset Arabic tweets

(c) Roc curve illustrating the performance of the ensemble Bert model on Test dataset Arabic tweets

Figure 2: Comparison of the proposed approach performance on on sub task A

Val set	Model	Sub-task A				
		FP	Recall	F-1 sarcastic	F1-macro	Acc
English	Baseline	**37.04**	**58.54**	**71.84**	**63.76**	**74.06**
	BERT Ensemble loss A	37.15	55.03	67.51	57 .00	72.33
	BERT Ensemble loss B	42.43	54.87	58.82	41.48	70.89
Arabic	Baseline	**23.60**	64.92	83.35	**77.75**	83.28
	BERT+ Ensemble loss A	25.61	**65.75**	**83.53**	77.41	**83.92**
	BERT Ensemble loss B	26.82	61.62	64.61	42.94	75.24

Table 1: Results of validation dataset for sub-task A, where BERT+ Ensemble loss A and BERT+ Ensemble loss B represents 80% /20% and 20% /80% combination of the cross entropy loss and negative log likelihood loss respectively. Where FP represents False positives

Test set	Model	Sub-task A				
		FP	Recall	F-1 sarcastic	F1-macro	Acc
English	Baseline	46.59	**52.66**	79.37	54.88	81.00
	BERT Ensemble loss A	**38.86**	52.25	**79.75**	**58.00**	80.07
	BERT Ensemble loss B	46.65	51.51	79.12	46.15	**85.71**
Arabic	Baseline	38.86	**52.25**	79.74	**57.99**	80.07
	BERT+ Ensemble loss A	**37.22**	51.28	76.10	56.60	73.79
	BERT+Ensemble loss B	40.13	52.61	**79.12**	46.15	**85.71**

Table 2: Results of Test dataset for sub-task A, where BERT+ Ensemble loss A and BERT+ Ensemble loss B represents 80% /20% and 20% /80% combination of the cross entropy loss and negative log likelihood loss respectively. Where FP represents False positives

F-1 sarcastic metric of 79.75% Its application on the Arabic language test set yielded an F-1 sarcastic metric of 79.12%, which was just below the value obtained from the baseline model which was 79.74%

Figures 1 and 2 show the Receiver Operating Characteristic (ROC) Curve of the different models applied to the test and validation data sets, which is a plot of the true positive rates against false-positive rates. It graphically shows and compares the performance of the baseline model and our ensemble loss model at all classification thresholds. It pictorially shows the evaluation of the strength of a model; with the bigger area under the curve indicating superior model performance and the smaller area under the curve signalling poorer model classification performance respectively.

5 Conclusion

This work has proposed the use of ensemble loss on the BERT model to detect whether a given sentence in English or Arabic language is sarcastic or not. The results indicate that the use of our ensemble loss on the Bert Model exhibits superior performance over the baseline models when applied to the English and Arabic test datasets. Future work will involve the undertaking of further investigation to find the optimal proportion of each combination that yields the optimal results. It will also involve the use of ensemble loss on Arabic-based Bert models such as Ara-Bert and other Bert-Based models to find their performance. Different types of losses will also be investigated and added to the ensemble loss portfolio in order to explore the different types with superior performance. Since the two language datasets used were unbalanced in nature, future work will also explore other data balancing techniques to help improve the results of our model.

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.

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *CoRR*, abs/1607.00976.

Adithya Avvaru, Sanath Vobilisetty, and Radhika Mamidi. 2020. Detecting Sarcasm in Conversation Context Using Transformer-Based Models. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 98–103, Online. Association for Computational Linguistics.

Imran Sarwar Bajwa and Muhammad Abbas Choudhary. 2006. A rule based system for speech language context understanding. *Journal of Donghua University (English Edition) Vol*, 23(06).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany. Association for Computational Linguistics.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection?

Akshay Khatri and Pranav P. 2020. Sarcasm detection in tweets with BERT and GloVe embeddings. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 56–60, Online. Association for Computational Linguistics.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.

Paul K Mandal and Rakeshkumar Mahto. 2019. Deep cnn-lstm with word embeddings for news headline sarcasm detection. In *16th International Conference on Information Technology-New Generations (ITNG 2019)*, pages 495–498. Springer.

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).

Bleau Moores and Vijay Mago. 2022. A survey on automated sarcasm detection on twitter.

Emmanuel Osei-Brefo, Thanet Markchom, and Huizhi Liang. 2021. UOR at SemEval-2021 task 12: On crowd annotations; learning with disagreements to

optimise crowd truth. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1303–1309, Online. Association for Computational Linguistics.

Krishna Parmar, Nivid Limbasiya, and Maulik V. Dhamecha. 2018. Feature based composite approach for sarcasm detection using mapreduce. *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, pages 587–591.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.