CS-UM6P at SemEval-2021 Task 7: Deep Multi-Task Learning Model for Detecting and Rating Humor and Offense

Kabil Essefar[1] **Abdellah El Mekki**[1] **Abdelkader El Mahdaouy**[1]
Nabil El Mamoun[2] **Ismail Berrada**[1]
[1]School of Computer Sciences, Mohammed VI Polytechnic University, Morocco
[2]Faculty of Sciences Dhar EL Mahraz, Sidi Mohamed Ben Abdellah University, Morocco
{firstname.lastname}@um6p.ma

Abstract

Humor detection has become a topic of interest for several research teams, especially those involved in socio-psychological studies, with the aim to detect the humor and the temper of a targeted population (e.g. a community, a city, a country, the employees of a given company). Most of the existing studies have formulated the humor detection problem as a binary classification task, whereas it revolves around learning the sense of humor by evaluating its different degrees. In this paper, we propose an end-to-end deep Multi-Task Learning (MTL) model to detect and rate humor and offense. It consists of a pre-trained transformer encoder and task-specific attention layers. The model is trained using MTL uncertainty loss weighting to adaptively combine all sub-tasks objective functions. Our MTL model tackles all sub-tasks of the SemEval-2021 Task-7 in one end-to-end deep learning system and shows very promising results.

1 Introduction

Humor is a human trait that defines the emotional and behavioral characteristics of an individual. It refers to the quality of being amusing, comic, sarcastic, etc. Most dictionaries define humor also as a message, whose ingenuity or verbal skill, or incongruity, that has the power to make individual laughing.

Humor and offensive language detection tasks are increasingly becoming hot research topics in Natural Language Processing (NLP) (Zampieri et al., 2019; Reyes et al., 2012; Gleason et al., 2019; van den Beukel and Aroyo, 2018; Cattle and Ma, 2018; Singh et al., 2020). Existing research works have tackled humor detection as either a binary classification problem (Weller and Seppi, 2019; Annamoradnejad, 2020) or a ranking task (Potash et al., 2017; Hossain et al., 2020; Zhang et al., 2019). Similarly, most research works on offensive language detection have proposed methods and approaches to discriminate between offensive and not-offensive texts (Zampieri et al., 2019, 2020), whereas, other research works have classified offensive content into more fine-grained levels (Wiegand et al., 2018; Kumar et al., 2018; Risch et al., 2020).

Fine-tuning pre-trained transformer-based language models on the target task data has shown state-of-the-art (SOTA) results in many NLP applications (Devlin et al., 2019; Liu et al., 2019). For instance, several research works on humor and offensive language detection have achieved SOTA performances using pre-trained transformer-based language models (Zampieri et al., 2019; Weller and Seppi, 2019; Zampieri et al., 2020).

In this paper, we describe our system submitted to the SemEval-2021 Task-7 (Sub-Tasks 1 and 2) (Meaney et al., 2021). We propose an end-to-end deep Multi-Task Learning (MTL) model based on RoBERTa Encoder (Liu et al., 2019) and task-specific attention layers. The attention mechanism is applied on top of the encoder's contextualized word embedding to extract task-specific features. The classification and regression modules are fed with their task-specific attention output and the shared pooled output of the encoder. In order to adaptively combine all tasks' losses, we employed the MTL uncertainty loss weighting method (Kendall et al., 2017). We also investigate the base and the large variants of BERT (Devlin et al., 2019) and RoBERTa encoders for both single-task and MTL. The obtained results show that our MTL model outperforms its single-task counterparts on both Task 1 and Task2. The best performances are obtained using RoBERTa-large encoder. Our system is ranked 18th, 9th, 7th and 20th on Sub-Tasks 1a, 1b, 1c and 2a, respectively.

The remainder of this paper is organized as fol-

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1135–1140
Bangkok, Thailand (online), August 5–6, 2021.

lows. Section 2 describes the SemEval-2021 task-7 and the provided data. Section 3 presents our MTL system. Section 4 summarizes the obtained results. Section 5 concludes the paper.

2 Task description

The SemEval-2021 Task 7 consists of two main tasks: the first task seeks recognizing and rating humor while the second task aims to rate offense (Meaney et al., 2021). To this end, the organizers have provided 8000 sentences for the training, and 1000 sentences for the validation and test. All training and validation sentences are labeled for humor detection and offense rating, while only humorous sentences are labeled for humor and controversy rating. The dataset is labeled by 20 annotators. They have a balanced set of age groups from 18 to 70.

2.1 Task 1: Humor detection

The aim is to predict four target values for the following sub-tasks:

- Task 1a: This sub-task is a binary classification task where the aim is to classify texts as humorous or not.
- Task 1b: This sub-task consists of predicting the humor degree of a text. The degree is based on the average rating (from 0 to 5) given by the annotators.
- Task 1c: This sub-task consists of predicting whether the humor rating would be considered controversial or not: i.e. whether or not the variance between the annotators' ratings is higher than the median rating.

2.2 Task 2: Offensive rating

This task has one sub-task for offense rating:

- Task 2a: This task predicts the degree of offense conveyed in a text regardless of its humor label. The offense degree varies from 0 (not offensive) to 5 (very offensive).

3 System description

We propose an end-to-end deep MTL model based on pre-trained transformer-based language model (Devlin et al., 2019; Liu et al., 2019) and task-specific attention layers. First, we apply the encoder to the input text in order to obtain its Contextual Word Embedding (CWE). The task-specific attention layers are applied on the CWE. The classifier (Task 1a, Task 1c) or the regressor (Task 1b, Task 2a) is fed with the concatenation of its task-specific attention output and the encoder's pooled output. The model is then trained to minimize the binary cross-entropy loss and the RMSE loss for the classification and regression tasks, respectively. Finally, these losses are combined using uncertainty loss weighting for MTL.

3.1 Transformer encoder

In order to recognize the most important patterns in an input text, we encode its using the state-of-art pre-trained transformer encoder. We compare four transformer encoders, namely BERT, BERT-Large, RoBERTa and RoBERTa-Large (Devlin et al., 2019; Liu et al., 2019).

The tokenizer of the encoder splits the input sentence into wordpeices $[T_1, T_2, ..., T_n]$ and encodes them using its vocabulary. The transformer encoder is fed with the encoded input and outputs the pooled embedding $h_{pooled} \in \mathbb{R}^{1\times d}$ (embedding of $[CLS]$ (resp. $< s >$) token of BERT (resp. roBBERTa)) and the CWE $H = [h_1, h_2, ..., h_n] \in \mathbb{R}^{n\times d}$ (d is the embedding dimension).

3.2 Task-specific attention layer

We use one task-specific attention layer for each task. Using H, the CWE of the input sentence, the attention mechanism (Bahdanau et al., 2015; Yang et al., 2016) extracts the task-specific representation s_* ($*$ denotes the task) as follows:

$$U = \tanh(W_a H))$$

$$\alpha = softmax(U^T W_\alpha)$$

$$s_* = \alpha \cdot H^T$$

where $W_a \in \mathbb{R}^{d\times 1}$ and $W_\alpha \in \mathbb{R}^{n\times n}$ are the trainable parameters of the attention layer, $U \in \mathbb{R}^{n\times 1}$ is the attention mechanism's context vector, and $\alpha \in [0, 1]^n$ weights $h_1, h_2, ..., h_n$ according to their contribution to the task objective.

3.3 Task Classification/Regression module

As the SemEval-2021 Task-7 consists of two classification tasks (1a and 1c) and two regression tasks (1b and 2), we employ two classification modules and two regression modules. Each of these task-specific module is composed of one hidden layer and one output layer, and takes as input the concatenation $[h_{pooled}, s_*]$ of the pooled output (h_{pooled}) and its task-specific attention output (s_*).

3.4 MTL objective

Our MTL model is trained to minimize the losses of the four tasks. Specifically, it minimizes the binary cross-entropy loss and the RMSE loss for classification and regression tasks, respectively. These losses are expressed as follows:

- Binary cross-entropy loss for humor classification

$$L_1(\hat{y}, y) = -\sum_{i=1}^{N}\sum_{j=1}^{2} y_i^j \log(\hat{y}_i^j)$$

- RMSE loss for Humor rating

$$L_2(\hat{y}, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_j)^2}$$

- Binary cross-entropy loss for Controversy classification

$$L_3(\hat{y}, y) = -\sum_{i=1}^{N}\sum_{j=1}^{2} y_i^j \log(\hat{y}_i^j)$$

- RMSE loss for offense rating

$$L_4(\hat{y}, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_j)^2}$$

where y and $\hat{y}$ are the ground-truth and the predicted values, respectively. In order to adaptively weight the losses of the four tasks, we combine them using MTL uncertainty loss weighting (Kendall et al., 2017), given by:

$$\begin{aligned} L_{total}(\hat{y}, y) &= \frac{1}{2\sigma_1^2}L_1(\hat{y}, y) + \frac{1}{2\sigma_2^2}L_2(\hat{y}, y) \\ &+ \frac{1}{2\sigma_3^2}L_3(\hat{y}, y) + \frac{1}{2\sigma_4^2}L_4(\hat{y}, y) \\ &+ \log(\sigma_1\sigma_2\sigma_3\sigma_4) \end{aligned}$$

where σ_i ($i = 1..4$) captures the amount of noise that exists in the output of each task, and used to tune the impact of each loss in MTL optimization. Finally, the MTL model is trained to minimize the overall loss L_{total} with respect to the network parameters as well as the noise parameters σ_i.

4 Results

4.1 Experiment Settings

We have evaluated the performance of our model and it single-task counterparts using both the *base* and the *large* models of BERT and RoBERTa:

- **BERT-base**: 12 transformer blocks, $d = 768$, 12 attention heads, and 110M parameters.
- **BERT-large**: 24 transformer blocks, $d = 1024$, 16 attention heads, and 336M parameters.
- **RoBERTa-base**: 12 transformer blocks, $d = 768$, 12 attention heads, and 125M parameters.
- **RoBERTa-large**: 24 transformer blocks, $d = 1024$, 16 attention heads, and 355M parameters.

For text preprocessing, we have implemented a simple pipeline that normalizes contractions. All evaluated models are trained using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1×10^{-5}. The batch size and the number of epochs are fixed to 16 and 5, respectively. We have investigated both single-task training and MTL for all tasks. It is worth mentioning that, for single-task learning, we also apply an attention layer on top of the contextualized word embedding. This has improved single-task models as well. All models are trained on the full train sets, validated on the validation set, and evaluated on the test set of each task. For evaluation purpose, we have used the shared task's evaluation metrics, namely the **F1-score**, the **Accuracy**, and the Root Mean Squared Error **RMSE**. It is worth mentioning that models' validation is preformed using the development set, while the presented results are obtained employing the test set.

4.2 Experiment Results

Table 1 presents the obtained results for all tasks using single-task and MTL models. The results show that our MTL model surpasses its single-task counterparts on all tasks. The large variants of BERT and RoBERTa encoders offer better performance compared to their base variants. The best performance is obtained using our MTL model on top of RoBERTa large encoder. These results can be explained by the fact that deep encoders can capture more complex pattern from the input text.

		Task 1a		Task 1c		Task 1b	Task 2
		Accuracy	F1-score	Accuracy	F1-score	RMSE	RMSE
Single-Task	BERT	0.9171	0.9318	0.4216	0.6023	0.5852	0.5649
	BERT-large	0.9372	0.9425	0.4296	0.6122	0.5748	0.5571
	RoBERTa	0.9408	0.9433	0.4302	0.6035	0.5711	0.5556
	RoBERTa-large	0.9422	0.9531	0.4415	0.6183	0.5688	0.5079
MTL	BERT	0.93	0.9361	0.4296	0.6032	0.5756	0.5511
	BERT-large	0.9421	0.9442	0.4316	0.6054	0.552	0.5533
	RoBERTa	0.945	0.9587	0.4423	0.6152	0.5578	0.5218
	RoBERTa-large‡	**0.951**	**0.9606**	**0.4537**	**0.6242**	**0.5401**	**0.4696**

Table 1: The obtained results using our MTL model and its single-task counterparts, with different encoders. The subscript ‡ denotes our official submission to SemEval-2021 Task-7. Results are obtained employing the test set.

	Task 1a		Task 1c		Task 1b	Task 2
	Accuracy	F-score	Accuracy	F-score	RMSE	RMSE
w/o task-attention	0.9335	0.9472	0.4456	0.6122	0.5636	0.4891
w/o uncertainty loss weighting	0.9498	0.9582	0.4516	0.6198	0.5516	0.4722
MTL RoBERTa-large	**0.951**	**0.9606**	**0.4537**	**0.6242**	**0.5401**	**0.4696**

Table 2: Ablation study of our MTL model using MTL RoBERTa-large encoder (w/o denotes without the corresponding component). Results are obtained using the test set.

Besides, MTL leverages useful signals from the related tasks.

To investigate the effectiveness of the task-specific attention layers and the uncertainty loss weighting on the performance of our MTL model, we have performed an ablation study. Table 2 presents the results of our model without these components. The results show that both components improve the performance of our MTL model. We achieve the most performance gain by incorporating the task-specific attention layers into our model. Besides, the adaptive losses weighting component outperforms the simple combination of task losses ($L_{total} = l_1 + l_2 + l_3 + l_4$).

5 Conclusion

In this paper, we have presented our system for humor and offense detection and rating. Our system consists of an end-to-end MTL model based on the state-of-art pre-trained transformer encoder and task-specific attention layers. The latter layers are applied on top of the contextualized word embedding to extract task-discriminative features. We have employed two classification and regression modules to tackle the four tasks. Our MTL model is trained to minimize the four tasks losses, while weighting them adaptively using the MTL uncertainty loss weighting. We have also investigated the performance of our MTL model as well as its single-task counterparts using four pre-trained transformer-based encoders. The best performances are obtained using our MTL model while employing RoBERTa-large encoder.

In future work, we would like to improve our model, by considering the relationship between the different tasks. Besides, we want to use our model not only to detect humorous and offensive content, but also to perform other related tasks.

References

Issa Annamoradnejad. 2020. Colbert: Using BERT sentence embedding for humor detection. *CoRR*, abs/2004.12765.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Sven van den Beukel and Lora Aroyo. 2018. Homonym detection for humor recognition in short text. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 286–291, Brussels,

Belgium. Association for Computational Linguistics.

Andrew Cattle and Xiaojuan Ma. 2018. Recognizing humour using word associations and humour anchor extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1849–1858, Santa Fe, New Mexico, USA. 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.

Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making memes accessible. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 367–376, New York, NY, USA. Association for Computing Machinery.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Semeval-2017 task 6:# hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowledge Engineering*, 74:1–12. Applications of Natural Language to Information Systems.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France. European Language Resources Association (ELRA).

Pranaydeep Singh, Nina Bauwelinck, and Els Lefever. 2020. Lt3 at semeval-2020 task 8 : multi-modal multi-task learning for memotion analysis. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval 2020)*, pages 1155–1162. Association for Computational Linguistics (ACL).

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing KONVENS 2018*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth*

Workshop on Semantic Evaluation, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Dongyu Zhang, Heting Zhang, Xikai Liu, LIN Hongfei, and Feng Xia. 2019. Telling the whole story: A manually annotated chinese dataset for the analysis of humor in jokes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6403–6408.