MilaNLP at WASSA 2021: Does BERT Feel Sad When You Cry?

Tommaso Fornaciari
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
tommaso.fornaciari@unibocconi.it

Federico Bianchi
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
f.bianchi@unibocconi.it

Debora Nozza
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
debora.nozza@unibocconi.it

Dirk Hovy
Bocconi University
Via Sarfatti 25, 20136
Milan, Italy
dirk.hovy@unibocconi.it

Abstract

The paper describes the MilaNLP team's submission (Bocconi University, Milan) in the WASSA 2021 Shared Task on Empathy Detection and Emotion Classification. We focus on Track 2 – Emotion Classification – which consists of predicting the emotion of reactions to English news stories at the essay-level. We test different models based on multi-task and multi-input frameworks. The goal was to better exploit all the correlated information given in the data set. We find, though, that empathy as an auxiliary task in multi-task learning and demographic attributes as additional input provide worse performance with respect to single-task learning. While the result is competitive in terms of the competition, our results suggest that emotion and empathy are not related tasks – at least for the purpose of prediction.

1 Introduction

Different researchers have been exploring emotion prediction from text (Abdul-Mageed and Ungar, 2017; Nozza et al., 2017). The WASSA-2021 shared task (Tafreshi et al., 2021) tackles the prediction of empathy (Track 1 of the challenge) and emotion (Track 2 of the challenge) in text. We, the MilaNLP lab, participated in Track 2 of the challenge. Nozza et al. (2020) show that Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) can provide accurate results for many different languages in different tasks. Indeed, we contributed to this year's WASSA workshop with two papers (Lamprinidis et al., 2021; Bianchi et al., 2021) that show that BERT can obtain good results in the emotion prediction task.

This system paper describes our approach to the emotion prediction *shared* task. Based on our previous experience with text classification tasks (Rashid et al., 2020; Fornaciari and Hovy, 2019; Uma et al., 2020), our initial idea was to use BERT, and support emotion prediction by adding an auxiliary task with the empathy score provided in the training set. Using such a Multi-Task Learning (MTL) setup can significantly boost performance on the main task, by exploring complementary information in the tasks, and by acting as a regularizer (a model that has to be able to predict more than one task is less prone to overfitting to any one of them). However, we unexpectedly find evidence for the opposite: using empathy as an auxiliary task in multi-task learning in this setting does *not* work as expected. In fact, adding empathy prediction hurts performance compared to a single-task model.

This finding adds to the literature that auxiliary tasks in MTL setups need to be related to the main task to help performance (Martínez Alonso and Plank, 2017). It also indicates that empathy is not directly a contributing factor to emotions, i.e., that there is no strong correlation between the two tasks.

2 Data

In this paper, we focused on emotion prediction (Track 2) of reactions to English news stories. The data set is an extended version of the one presented in Buechel et al. (2018). Each instance corresponds to an empathic reaction to news stories extracted by popular online news platforms. A set of 1860 training documents annotated with seven emotions was given (see Table 2 for the data set size). With each text document, an empathy score that ranges from 1 to seven has been associated; in Table 1 we show some examples of text with the emotion and the empathy that come from the data set.

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 269–273

Text	Emotion	Empathy
it is really diheartening to read about these immigrants from this article who drowned. it makes me feel anxious and upset how the whole ordeal happened. it is a terrible occurrence that this had to happen at the mediterranean sea. thankfully there were some survivors. the fact that babies were lost makes it that much more emotional to read all of this	*sadness*	*5.667*
This is a crazy story with so many facets to it, omg. I mean on one hand, I don't support an eye for an eye. I don't support the death penalty and I don't support blinding someone. BUT on the other hand, this is a country where women really struggle and the justice system is not well developed. ALSO, he blinded a FOUR YEAR OLD GIRL. What the fuck is wrong with this guy. So if this was in America I would not support it, but I don't feel right condemning the actions of an entirely different country for doing what they felt needed to be done.	*anger*	*1*

Table 1: Examples of documents with the emotion and the empathy that come from the data set. One document has relatively high empathy, while the second one has very low empathy.

Emotion	sadness	neutral	fear	anger	disgust	surprise	joy
Training Set	647	275	194	349	149	164	82
Development Set	98	31	76	25	12	14	14

Table 2: Training and Development set sizes

3 System Description

In this section, we describe the different configurations of the system we use for the emotion prediction task. We remind that we focused only on Track 2 of the WASSA Shared Task challenge.

3.1 Experimental Conditions

The data set allows building both Multi-Input (MI) and Multi-Task Learning (MTL) models. We tried both methods, separately and together, and we compare them with a single-input, single-task model. The single-input, single-output model uses text as input and predicts emotions.

We create three MI models. All of them take the texts as input: this is our Single-Task Learning (STL) model. We also build three Multi-Input (MI) models where, besides the text, we also include gender information (2-input model, MI1), gender and income (3-input model, MI2), and gender, income, and Interpersonal Reactivity Index (IRI) (4-input model, MI3).

Given the availability of further dependent variables, we create a Multi-Task Learning (MTL) model that takes the text as only input and jointly predicts emotions (classification task with categorical cross-entropy), empathy, and distress (regression task) (MTL2).

Lastly, we implement an MI-MTL model that exploits text, gender, income, and IRI as input and predicts emotions, empathy, and distress (MI3-MTL2).

3.2 Architectures of the Models

In all our models, we use the BERT language model (Devlin et al., 2019). In particular, we use the `bert-large-uncased` model for English, that is made of 336M parameters. The model comes with its own tokenizer that we use to extract a word $\times$ contextual embedding matrix for each text. We use such matrix as input for a single-layer, single-head Transformer, following Vaswani et al. (2017), that is in charge to detect specific patterns of emotion. Lastly, a fully connected layer provides the output prediction.

In the MTL models, we have a separate fully connected layer for each task. Even though the different tasks concern the prediction of values being in

	Acc	P	R	F1
STL	**58.48**	**54.64**	**47.05**	**48.65**
MI1	52.19	50.23	36.63	36.69
MI2	44.76	32.34	29.11	26.09
MI3	56.19	41.27	39.15	38.31
MTL2	51.43	49.53	38.41	38.96
MI3-MTL2	34.67	29.03	19.09	14.19

Table 3: Accuracy (Acc), Precision (P), Recall (R) and F1-score (F1) on the Development set. Significance levels over STL: $^*: p \leq 0.05$

a similar scale, we also tried to add a normalization layer, with the aim of keeping such scale similar for all the tasks. We did not find performance improvements, therefore we show the results without normalization.

In the MI models, besides the BERT representation, we also use vectors of size 3, 1, and 4 for gender categories, income, and IRI values respectively. The gender vectors are one-hot encoded; income and IRI values are (column-wise) normalized float values.

As loss functions, we use cross-entropy for the classification, and mean squared error for the two regression tasks. We use Adam optimizer (Kingma and Ba, 2015). We select the models through an early-stopping that requires a decrement rate on the development set's loss lower than 12% for three consecutive epochs. Our learning rate is 0.002, drop-out probability 0.2, and batch size 64, manually tuned.

To test the significance of the possible improvements over the STL base model, we use a bootstrap sampling test (Søgaard et al., 2014), with 1000 loops and a sample size of 30%.

4 Discussion

Table 3 presents our results. In many cases (Ruder, 2017), using multi-task learning on related tasks can help performance, especially when labeled data is sparse or unbalanced. Intuitively, it would seem that empathy and emotion would make for good candidates. However, in our experiments with this combination, we found a negative effect of empathy on emotion prediction. Upon closer inspection, that makes sense: being empathetic towards someone does not necessarily entail a particular emotion.

Somewhat surprisingly, neither do demographic attributes. Various prior works have found those factors to help in classification settings (Volkova et al., 2013; Hovy, 2015; Lynn et al., 2017), especially with MTL (Ruder, 2017; Benton et al., 2017; Li et al., 2018). However, they do not seem to improve emotion classification here.

Therefore, for the shared task submission we choose the STL model, which showed the highest F-measure on the development set in our experimental conditions. On the test set, we obtained an F1-score equal to 48.6; with this score our team ranked third in the Track 2.

5 Conclusion

Our results seem to suggest that the combination between empathy and emotion is a difficult task. Given our low scores in the multi-task setting we also speculate on the fact that the two tasks might not be so easy to relate. Future work should consider better ways to aggregate the information coming from these two models.

6 Acknowledgments

The authors are members of the Data and Marketing Insights research unit in the Bocconi Institute for Data Science and Analysis.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2021. Feel-it: Emotion and sentiment classification for the italian language. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. 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.

Tommaso Fornaciari and Dirk Hovy. 2019. Geolocation with Attention-Based Multitask Learning Models. In *Proceedings of the 5th Workshop on Noisy User-generated Text (WNUT)*.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

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

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 44–53, Valencia, Spain. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Debora Nozza, Elisabetta Fersini, and Enza Messina. 2017. A multi-view sentiment corpus. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 273–280, Valencia, Spain. Association for Computational Linguistics.

Farzana Rashid, Tommaso Fornaciari, Dirk Hovy, Eduardo Blanco, and Fernando Vega-Redondo. 2020. Helpful or hierarchical? predicting the communicative strategies of chat participants, and their impact on success. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2366–2371.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso. 2014. What's in a p-value in nlp? In *Proceedings of the eighteenth conference on computational natural language learning*, pages 1–10.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.