AILAB-Udine@SMM4H'22: Limits of Transformers and BERT Ensembles

Beatrice Portelli
portelli.beatrice@spes.uniud.it
University of Udine, Italy
University of Naples Federico II , Italy

Simone Scaboro
scaboro.simone@spes.uniud.it
University of Udine, Italy

Emmanuele Chersoni
emmanuele.chersoni@polyu.edu.hk
The Hong Kong Polytechnic
University, Hong Kong

Enrico Santus
esantus@gmail.com
DSIG - Bayer Pharmaceuticals, New Jersey, USA

Giuseppe Serra
giuseppe.serra@uniud.it
University of Udine, Italy

Abstract

This paper describes the models developed by the AILAB-Udine team for the SMM4H'22 Shared Task. We explored the limits of Transformer based models on text classification, entity extraction and entity normalization, tackling Tasks 1, 2, 5, 6 and 10. The main takeaways we got from participating in different tasks are: the overwhelming positive effects of combining different architectures when using ensemble learning, and the great potential of generative models for term normalization.

1 Introduction

Transformer-based models are the backbone of state-of-the-art solutions for a lot of NLP tasks. The real strength of these models (like BERT Devlin et al., 2018, GPT Radford et al., 2019, and their variants Joshi et al., 2019; Gu et al., 2020) stands in the pre-training phase which permits them to have extensive language knowledge. This is particularly helpful in tasks where the amount of training data is restricted, like the ones addressed in this workshop (Weissenbacher et al., 2022).

In this work we used a variety of pretrained Transformers models for the tasks. We refer to Table 1 for a summary of their names in the Huggingface library and the shorthand version of their name used in this report.

Short name	Model name in the Huggingface library	Reference
GPT-2	gpt2	(Radford et al., 2019)
BERT$_{\mathrm{Eng}}$	bert-base-uncased	(Devlin et al., 2018)
BERT$_{\mathrm{Mul}}$	bert-base-multilingual-uncased	(Devlin et al., 2018)
BERT$_{\mathrm{Med}}$	microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract	(Gu et al., 2020)
BERT$_{\mathrm{Span}}$	SpanBERT/spanbert-base-cased	(Joshi et al., 2019)
RoBERTa$_{\mathrm{Twi}}$	cardiffnlp/twitter-roberta-base-sentiment	(Barbieri et al., 2020)
RoBERTa$_{\mathrm{XML}}$	xlm-roberta-base	(Conneau et al., 2019)

Table 1: List of pretrained models used for the tasks and the abbreviations used in this report.

2 Simple Classification (Task 5 / 6)

Task 5 and 6 both entail the simple classification of tweets (binary or ternary) in a class-unbalanced setting. Task 5 consists in the ternary classification of Spanish tweets about COVID-19 symptoms as: containing literature/news reports (News, 60%), containing personal reports (Pers, 16%) or reporting about someone else's symptoms (Non-Pers, 24%). Task 6 consists in the binary classification of English tweets regarding COVID-19 vaccinations as: general vaccine chatter (Chatter, 89%) or personal reports confirming the vaccination status of the user (Pers, 11%). For both tasks the text preprocessing consisted in replacing all usernames with "user" and all URLs with "(see url)" or "(ver url)" ("(see url)" in Spanish).

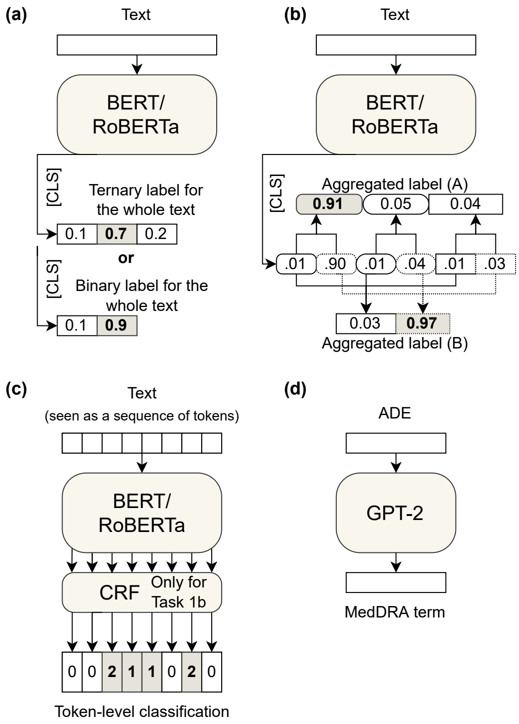

Figure 1: Summary of the model architectures used in the different tasks.

Proceedings of the 29th International Conference on Computational Linguistic, pages 130–134
October 12–17, 2022.

2.1 Models

We tackled the tasks using a simple Transformer-based model with a classification head, that is a linear layer applied to the output embedding of the [CLS] token (see Figure 1a). This layer maps the embedding to either two or three classes depending on the task. The kind of pretrained model was chosen based on the characteristics of the input text, such as language. For each task we selected two kinds of pretrained models with different characteristics to try and combine them, and ensembled their predictions. The selected models were: two BERT_{Mul} and one $\text{RoBERTa}_{\text{XML}}$ for Task 5, two BERT_{Mul} and one BERT_{Eng} for Task 6. They were chosen by training and evaluating 5 models of each kind locally on a 70-30 random split of the training data, and selecting the ones with the higher performance on their respective test fold. In Task 5, BERT_{Mul} models were trained for 10 epochs while $\text{RoBERTa}_{\text{XML}}$ models for 15 epochs. In Task 6, BERT_{Mul} models were trained for 5 epochs while BERT_{Eng} models for 4 epochs. The final label for the task was chosen via majority vote.

2.2 Results

Table 2 shows the results of the base models and their ensemble for both tasks on the validation set.

Task	Model	P	R	F1
5	BERT_{Mul} (1)	0.826	0.826	0.826
5	BERT_{Mul} (2)	0.833	0.833	0.833
5	$\text{RoBERTa}_{\text{XML}}$	0.823	0.823	0.823
5	**Ensemble**	**0.838**	**0.838**	**0.838**
6	BERT_{Mul} (1)	0.875	0.734	0.799
6	BERT_{Mul} (2)	0.946	**0.748**	**0.835**
6	BERT_{Eng}	0.928	0.715	0.807
6	**Ensemble**	**0.954**	0.741	0.834

Table 2: Task 5 and Task 6 results on the validation set.

Ensembling different model typologies had a positive effect on Task 5, as the overall performance is higher than any model on its own. Looking at the confusion matrices in Figure 2, we see that most of the improvements come from a better accuracy in classifying Pers samples and distinguishing them from Non-Pers ones. The precision on Pers class goes from 0.68 (single models) to 0.72 (Ensemble) and the percentage of Pers samples classified as Non-Pers lowers from 0.27 to 0.23.

As regards Task 6, the Ensemble has a higher precision than each individual model, but a lower recall. BERT_{Mul} (2) had significantly higher metrics compared to the other two models, and the majority vote might have favored the most frequent (incorrect) prediction of the other two models. Looking at the confusion matrices in Figure 3, we can see that the Ensemble model has a higher precision on the most frequent class (Chatter) compared to the single models, but the two weaker models severely hampered the performance of BERT_{Mul} (2).

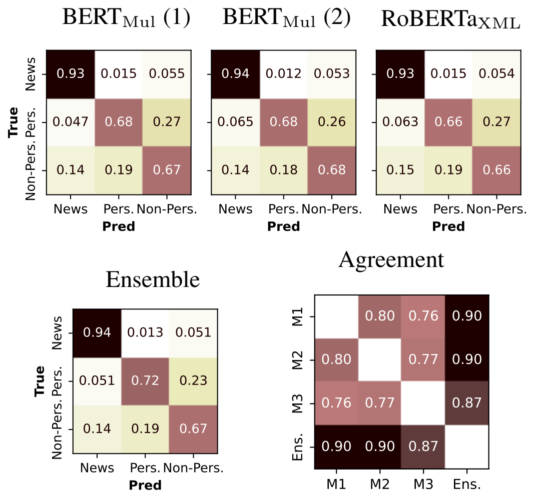

Figure 2: Task 5. Confusion matrices for the three separate models and their ensemble, and agreement matrix.

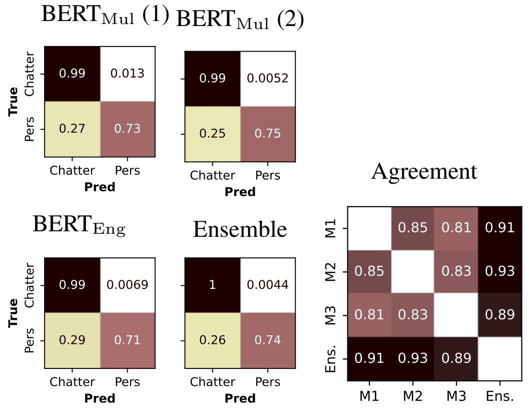

Figure 3: Task 6. Confusion matrices for the three separate models and their ensemble, and agreement matrix.

The main differences between the models ensembled for Task 5 and Task 6 is that the models used for Task 5 had different base architectures (BERT vs RoBERTa), while the ones for Task 6 were all based on BERT. If we calculate the agreements between the models using Cohen's Kappa Cohen, 1960, we see that the models used for Task 6 had higher agreement than the ones in Task 5 (compare the agreement matrices in Figures 2 and 3). The lower agreement for Task 5 is likely caused by the use of different model architectures, and it might lead to higher performance when combining the predictions.

3 Multitask Classification (Task 2a+2b)

Task 2a and 2b (Davydova and Tutubalina, 2022) consist in the classification of English tweets containing opinions about mandates during the COVID-19 pandemic. The tweets can deal with three topics: Face Masks (M), Stay At Home Orders (H) and School Closures (S). Task 2a is a ternary stance classification (Against, None, Favor), while Task 2b is a binary premise classification (1, 0) to determine whether the tweet is argumentative or not.

3.1 Models

We use the same architecture to solve both tasks, reformulating them as a single 6-way classification with labels: Against-1, Against-0, None-1, None-0, Favor-1 and Favor-0. We use a simple Transformer-based model with a classification head on top of the [CLS] embedding. The output of the classification head is a probability distribution over the six labels. At inference time, the probabilities are aggregated in three or two classes according to the task (e.g., Against=Against-1 + Against-0 for Task 2a or 1 = Against-1 + None-1 + Favor-1 for Task 2b). This process is illustrated in Figure 1b.

The text preprocessing is the same as Task 6. The input for the models was formatted as “About CLAIM. [SEP] TWEET_TEXT”, where CLAIM is one of the three topics. We finetuned a BERT$_{\text{Eng}}$ model for 4 epochs on all training data. To increase the robustness of the model for Task 2a, we also finetuned two RoBERTa$_{\text{Twi}}$ models for 5 epochs on the three-way classification Against/None/Favor, leveraging the model’s pretrained weights for sentiment classification on Twitter. We then ensemble the predictions of the two RoBERTa$_{\text{Twi}}$ models and the BERT$_{\text{Eng}}$ model for Task 2a (similarly to Task 5/6).

3.2 Results

Table 3 reports the metrics for Task 2a and 2b on the validation set. The Ensemble for Task 2a achieves higher metrics compared to the single models in two out of the three topics (M and H), as well as on the overall F1 score. The F1 score of the single models differ up to 7-9 points between each other, yet their interaction leads to a higher score overall. Figure 4 shows the agreement between the three models, which is even lower than the one recorded for Task 5. This further strengthens the hypothesis that using different architectures and models with high disagreements leads to an ensemble with higher performance.

Task	Model	F1$_M$	F1$_S$	F1$_H$	F1
2a	RoBERTa$_{\text{Twi}}$ (1)	0.840	0.677	0.819	0.779
2a	RoBERTa$_{\text{Twi}}$ (2)	0.816	0.700	**0.821**	0.779
2a	BERT$_{\text{Eng}}$	0.749	0.610	0.742	0.700
2a	**Ensemble**	**0.855**	**0.722**	0.807	**0.795**
2b	**BERT$_{\text{Eng}}$**	**0.786**	**0.818**	**0.814**	**0.806**

Table 3: Task 2a and 2b results on the validation set.

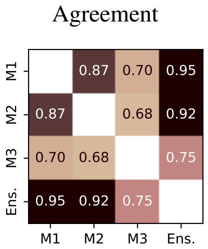

Figure 4: Task 2a. Agreement matrix for the three models and their ensemble.

4 Disease Extraction (Task 10)

Task 10 (Gasco et al., 2022) consists in extracting disease mentions from Spanish tweets.

4.1 Models

We solved this task using a simple Transformer-based model with a token-classification head, that is a linear layer applied to the output embedding of each token (see Figure 1c) with three output classes. These represent the BIO tagging scheme (Begin-Inside-Outside), commonly used to mark the presence of Named Entities in NER tasks. This straightforward method has been previously used in SMM4H Tasks for ADE extraction (Portelli et al., 2022), and we were interested in testing if it was possible to adapt it to Disease Extraction with minimal changes. The text preprocessing is the same as Task 5. We used a BERT$_{\text{Mul}}$ model (with multilingual pretraining) and trained it for 5 epochs on the training data of SocialDisNER without using additional resources.

4.2 Results

Table 4 reports the strict and relaxed metrics on the validation set. There is a gap of 40 points between the two, which is almost double what is usually reported in ADE extraction tasks. This means that the model was able to identify the broad area of text containing the disease, but not to pinpoint it. This goes to show that a Disease extraction system needs more mechanisms in place to precisely extract the relevant text.

Task	Model	P	R	F1
10	BERT$_{\text{Mul}}$ (Strict)	0.543	0.498	0.520
10	BERT$_{\text{Mul}}$ (Relaxed)	0.946	0.865	0.904

Table 4: Task 10 results on the validation set.

5 ADE Normalization (Task 1c)

Task 1c consists in mapping ADE mentions from English tweets to their corresponding MedDRA terms (formal medical terms). The dataset (Magge et al., 2021) was developed with three sub-tasks in mind, so to perform Task 1c it is necessary to complete the two preliminary Tasks 1a (binary classification ADE/noADE) and 1b (ADE extraction).

5.1 Models

Task 1a and Task 1b were not the main focus of our work, so we tackled them with simple and effective strategies seen in other tasks. Task 1a was solved with a BERT$_{\text{Med}}$ model with a binary classification head, trained as seen in Task 6, without model ensembling. For Task 1b we used a model previously developed for the same task the SMM4H'19 Shared Task Portelli et al. (2021, 2022). It consists of a BERT$_{\text{Span}}$ for token classification (see model for Task 10) combined with a Conditional Random Field (CRF) module (see Figure 1c).

For Task 1c, we used a GPT-2 model, trained to take as input a ADE and generate the string corresponding to the correct MedDRA term (e.g., "feel like crap" → "malaise"). GPT-2 was trained on the whole training set for 15 epochs.

5.2 Results

Table 5 reports the results of the models on the validation set. The models for Tasks 1a and 1b achieve average performance. We report the metrics for Task 1c in two ways: calculating them on the output of BERT$_{\text{Span}}$ (same procedures used on the blind test set); and using an oracle for Task 1b (that is, giving as input to GPT-2 only the correct ADEs). Using the oracle, we see that the model developed for Task 1c has a very high accuracy (0.759) if given the correct ADEs. Applying GPT-2 to the predictions of BERT$_{\text{Span}}$ leads to lower metrics due to the low quality of the preceding steps. The proposed model for Task 1c also performed extremely well on the blind test set, where it achieved results well over the average despite the low performance reached on Task 1b (see Table 6).

Task	Model	P	R	F1
1a	BERT$_{\text{Med}}$	0.663	0.477	0.544
1b	BERT$_{\text{Span}}$	0.295	0.851	0.438
1c	GPT-2 (from BERT$_{\text{Span}}$)	0.219	0.632	0.325
1c	GPT-2 (from oracle)	0.759	0.759	0.759

Table 5: Task 1 results on the validation set.

6 Results on the Test Set

The following table reports the metrics of all presented models on the test set, together with the reference scores supplied by organizers. Models were *not* re-trained using validation data, with the exception of Tasks 1a and 1b.

		Strict			Relaxed		
Task	**Model**	**P**	**R**	**F1**	**P**	**R**	**F1**
1a	Our	.607	.386	.472			
1a	Average	**.646**	**.497**	**.562**			
1b	Our	**.360**	.254	.298	.489	.344	.404
1b	Average	.344	**.339**	**.341**	**.539**	**.517**	**.527**
1c	Our	**.243**	**.171**	**.201**	**.294**	**.207**	**.243**
1c	Average	.085	.082	.083	.120	.112	.116
2a	Our			.529			
2a	Average			.491			
2a	Median			**.550**			
2b	Our			**.649**			
2b	Average			.574			
2b	Median			.647			
5	Our	.840	.840	.840			
5	Median	.840	.840	.840			
5	Baseline	**.900**	**.900**	**.900**			
6	Our	**.930**	.750	**.830**			
6	Median	.900	.680	.770			
6	Baseline	.900	**.770**	**.830**			
10	Our	.504	.461	.481			
10	Average	.680	.677	.675			
10	Median	**.758**	**.780**	**.761**			

Table 6: Results for all tasks on the blind test set.

7 Conclusions

We explored the use of simple Transformer-based architectures for several tasks proposed by SMM4H'22. The most noticeable phenomena we encountered were: the collaborative effect of different architectures (e.g., BERT and RoBERTa) when used in ensemble learning (Task 2 and 5, as opposed to Task 6); the efficacy of generative models for term normalization (Task 1c); and the low transferability of methods developed for ADE extraction to Disease detection (Task 10).

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.

Vera Davydova and Elena Tutubalina. 2022. Smm4h 2022 task 2: Dataset for stance and premise detection in tweets about health mandates related to covid-19. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, and Martin Krallinger. 2022. Overview of the SocialDisNER shared task on detection of diseases mentions from healthcare related and patient generated social media content: methods, evaluation and corpora. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. 2021. Bert prescriptions to avoid unwanted headaches: a comparison of transformer architectures for adverse drug event detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1740–1747.

Beatrice Portelli, Daniele Passabì, Edoardo Lenzi, Giuseppe Serra, Enrico Santus, and Emmanuele Chersoni. 2022. *Improving Adverse Drug Event Extraction with SpanBERT on Different Text Typologies*, pages 87–99. Springer International Publishing, Cham.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications #smm4h shared tasks at coling 2022. In *Proceedings of the Seventh Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages –.