Fraunhofer FKIE @ SMM4H 2022: System Description for Shared Tasks 2, 4 and 9

Daniel Claeser
Fraunhofer FKIE
Fraunhoferstraße 20
53343 Wachtberg
daniel.claeser@fkie.fraunhofer.de

Samantha Kent
Fraunhofer FKIE
Fraunhoferstraße 20
53343 Wachtberg
samantha.kent@fkie.fraunhofer.de

Abstract

We present our results for the shared tasks 2, 4 and 9 of the SMM4H Workshop at COLING 2022 achieved by succesfully fine-tuning pre-trained language models on the downstream tasks. We identify the occurence of code-switching in the test data for task 2 as a possible source of considerable performance degradation on the test set scores. We successfully exploit structural linguistic similarities in the datasets of tasks 4 and 9 for training on joined datasets, scoring first in task 9 and on par with SOTA in task 4.

1 Introduction

This contribution describes the system submissions for thee shared tasks at the Social Media Mining for Health (#SMM4H) Workshop at COLING 2022 (Weissenbacher et al., 2022). We participated in tasks 2, 4 and 9.

All models were developed using the Flair framework (Akbik et al., 2019) and we used the pre-trained models based on PyTorch (Paszke et al., 2019) provided by Huggingface (Wolf et al., 2020). Based on previous experience and the baseline results provided for the stance detection COVID-19 tweets in (Glandt et al., 2021), we focused mainly on the models BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and BERTweet (Nguyen et al., 2020).

2 Task 2

We participated in both subtasks of shared task 2. In task 2a, stance detection, participants were asked to determine an author's stance in relation to three different mandates related to the COVID-19 pandemic (Davydova and Tutubalina, 2022). The following tweets are examples relating to the mandate school closures:

- (FAVOR) - @anonymous NO TO REOPEN SCHOOLS.
- (AGAINST) - Society is bound to fall If Schools fall.
- (NONE) - The UK government tried to reopen schools and the people of Scotland refused.

In task 2b, premise classification, the aim is to determine whether or not a tweet contains an argument that could be used to convince an opponent about one of the given COVID-19 mandates. It is a binary classification task in which the data is annotated as positive (contains a premise) or negative (does not contain a premise). The following two tweets are examples from the mask wearing mandate:

- (1) - If masks work, then why are people working from home?
- (0) - @Anonymous @Anonymous is this about mask wearing?

The training data consists of 3,556 tweets, the validation data of 600 tweets, and the test data of 10,000 tweets.

2.1 System Description

We conducted preliminary experiments using the validation data as a held-out test set. The preliminary models were submitted during the evaluation phase of the shared task. For training data, we split the original training data into a train (3,000 tweets) and development (556 tweets) set. For this task, we used RoBERTa-large and BERTweet-large transformer document embeddings. RoBERTa-large is based on the BERT-large architecture, has 24-layers, 1,024 hidden layer dimension and 16 attention heads with 335M parameters. BERTweet-large is a language model pre-trained specifically on 850M English tweets, 5 million specifically related to the COVID pandemic, and is in turn based on the RoBERTa training procedure and has the same architecture.

Proceedings of the 29th International Conference on Computational Linguistic, pages 103–107
October 12–17, 2022.

Model	Weighted F1
RoBERTa	0.7458 - 0.7856
BERTweet	0.6889 - 0.7804

Table 1: Eight-fold cross validation for stance detection using RoBERTa and BERTweet.

Model	Weighted F1
RoBERTa	0.7769 - 0.8224
BERTweet	0.7822 - 0.8138

Table 2: Eight-fold cross validation for premise classification using RoBERTa and BERTweet.

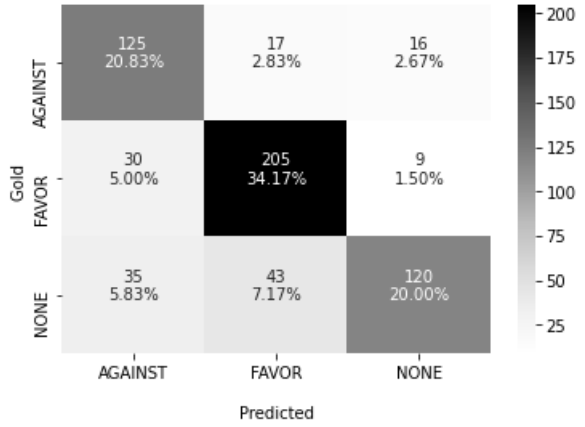

Figure 1: A confusion matrix showing the errors made by the stance detection system.

More specifically after parameter optimization, we fine-tuned both the RoBERTa-large and BERTweet large models and trained the embeddings using a learning rate of 0.005, with a mini-batch size of 32, for a maximum of 50 epochs for the stance detection task. For the premise classification sub-task we found similar parameters to be optimal, except we changed the mini-batch size to 16.

2.1.1 Stance Detection

Table 1 shows the result of performing eight-fold cross-validation at the training stage for the two different pre-trained models. The highest and the lowest weighted F1 scores are shown in the table, with the RoBERTa model slightly outperforming the BERTweet model with an F1 of 0.7856.

We conducted an error analysis on the misclassifications made by the systems. The confusion matrix in Figure 1 shows that the most errors are made in the class 'NONE'.

2.1.2 Premise Classification

We followed a similar procedure for premise classification, the results of which can be found in Table 2. Both models perform similarly, with RoBERTa (weighted F1 0.8224) slightly outperforming BERTweet (weighted F1 0.8138), and the results show that the variance in the eight system runs is 4.55 for RoBERTa and 3.16 for BERTweet.

2.2 Final Results

Comparing the results for stance detection to those in a paper by (Glandt et al., 2021), we wanted to maximize the available training data in order to boost performance of our systems. To train the final models, we incorporated the original validation data as training data for the final models. The data was split as follows: train (3,500), development (556) and test (10,000). The test data was pre-processed to match the training data using the Python emoji [1] library.

To gain a better understanding of the differences in scores, we examined the three different datasets provided by the task organizers. Specifically, we observed that the test data contains many more multilingual tweets compared to the training and validation data (see Table 3). In total, 2.93% of test tweets are identified as being non-English, using the FastText language identification model (Joulin et al., 2016). Furthermore, 13.2% of tweets receive a confidence score of less than 0.4 for English, indicating that the language is not so clearly identifiable. An example of a tweet containing Spanish-English code-switching can be found below (Poplack, 1980). Since the pre-trained language models are monolingual and have been fine-tuned on monolingual data, it is imaginable that the models struggle with accurately classifying the test data. It would be interesting to further analyze this issue and other reasons for inaccuracies once the gold standard for the test data is released.

- —échale un vistazo a esto… … a fair piece on comprehensive contrasting views on the virus so-called 'crisis'…

3 Task 4

Task 4 is a binary classification task in which participants were asked to determine whether a tweet contains a self-report of an exact age or not. For example, in tweet 1 below, the person posting the

[1] https://pypi.org/project/emoji/

#Tweets	Train	Val.	Test
Total	3556	600	10.000
English	3548	599	9707
Other	8	1	293
conf.< 0.4	293	48	1316

Table 3: Distribution of the languages of tweets in the data. Conf. refers to the confidence score given by the language detection algorithm. The table includes all tweets with a confidence score lower than 0.4 out of 1.

tweet reports that they are currently 29. In tweet 0, an age is reported, but it is annotated as negative because the age that is reported is that of the user's dog, and not their own age.

- (1) - My birthday is in 9 days. And just am here to say I am enjoying being still 29 the fullest.
- (0) - My 15yo Lab got into an unopened box of breakfast cereal, which made him sick...: I guess you can't feed an old dog new Trix.

The data consists of 8,800 tweets as training data, 2,200 tweets as validation data, and 10,000 tweets as test data.

3.1 System Description

We trained various configurations of BERT-large and RoBERTa-large (see section 2.1). As a specific contribution, we took advantage of the expected limited inventory available to authors independent of medium or register to express, both explicitely and implicitely, age: A basic n-gram analysis showed a high degree of overlap between the datasets of tasks 4 and 9 despite the differences in medium, register, topic and document length. Trigrams explicitely specifying age like 'n years old', '[the] age of n [years]' as well as trigrams enabling deduction of age in context, such as 'n years ago' are similarly distributed in both datasets to a degree that motivated the idea of merging both training sets.

We found the augmented model to clearly outperform the best configuration trained on the task 4 specific dataset only: Providing the model with an additional 9,000 training instances from the Reddit dataset led to an improved F1 of 0.9526, a statistically significant (8x cross-validated, p=0.000046) margin of 0.0193 over the best-performing model employing only the task-specific training set (F1 0.9333). The best results were achieved within 10 epochs at a learning rate of 0.005, AdamW optimizer (Loshchilov and Hutter, 2017) and a mini-batch size of 8. In order to retain good generalization capability in light of unknown label distribution in the blind test set, we evaluated for macro-F1 during fine-tuning.

Model	Macro F1	SD
RoBERTa-large	0.9333	0.0069
RoBERTa-large+T9	0.9526	0.0041

Table 4: Eight-fold cross validation for classification of tweets self-reporting exact age with basic and augmented training data.

3.2 Error Analysis

Eight cross-validation runs with models trained in the best-performing parameter configuration produced 913 misclassification instances out of 17,600 samples (551 false positives and 362 false negatives composed of multiple occurences of 148 and 94 unique instances, respectively). The mean error margin was 0.051875 (spread 107-126/2200, σ=6.07), with an average false positive rate of 0.0313 (51-87, σ=11.68) and a false negative rate of an average 0.0206 (36-57, σ=8.2424).

4 Task 9

Task 9 is very similar to task 4, in that it also involves a binary classification task on self-reported ages. The main difference is that the data stems from Reddit, not Twitter, and that the data is disease-specific and was collected using specific keywords related to dry eye disease. For example, in (1) a specific age is provided, and the second post in annotated as (0) because the user only provides an age range.

- (1) - How old are you? I would be surprised if your eyes have stabilized after only 5 years unless you were in your 30s when diagnosed. I was diagnosed at 21, had CXL in left eye that year and been monitoring both eyes since (now 33). Still waiting fir them to stabilize...
- (0) - is a .5D reduction in astigmatism for a developing cataract (age 60's) likely to be due to changes in the lens or cornea ?

The training data consists of 9,000 posts, the validation data of 1,000 posts, and the test data of 2,000 posts.

Model	Macro F1	SD
RoBERTa-large	0.9520	0.014043
RoBERTa-large+T4	0.9695	0.008629

Table 5: Eight-fold cross validation for classification of Reddit posts self-reporting exact age with basic and augmented training data.

4.1 System Description

In light of the performance gains on the evaluation set of task 4 from training a model based on the combined training sets of tasks 4 and 9, we opted to apply the same strategy for task 9 with its identical binary classification goal: Augmenting the tasks' original training data by the dataset provided for training task 4. Providing the model with an additional 9,000 training instances from the Twitter dataset led to an improved validation set F1 of 0.9695, a statistically significant (8x cross-validated, p=0.0033173) margin of 0.0175 over the best-performing model employing only the task-specific training set (F1 0.9520). Unsurprisingly, this result was accomplished by a configuration identical to that of task 4.

4.2 Error analysis

Eight cross-validation runs with models trained in the best-performing parameter configuration yielded 206 misclassification instances out of 8,000 samples (118 false positives and 88 false negatives composed of multiple occurences of 35 and 22 unique instances, respectively). The mean error margin was 0.02575 (spread 22-30/1000, σ=2.39), with an average false positive rate of 0.01475 (11-22, σ=3.38) and a false negative rate of an average 0.011 (6-16, σ=2.78).

Eleven unique samples accounting for 53 instances of false positive classification should actually be considered true positive since their gold label apparently did not comply to the annotation guidelines. Six of those samples contained an explicit mention of an exact age and five contained implicit, abbreviated or indirect (inferable) age mentions. Unambiguous examples of this phenomenon are e.g.

- (11287) [...] I am now 66 years old, finished high school, college, [...]
- (11239) [...] I was 26 when I was diagnosed too (am now 28) [...]

The complete list of such instances in the conducted validation runs is *11012, 11153, 11173, 11396, 11833, 11294, 11863, 11539* with more examples supposedly identifiable with additional runs.

Considering these samples to be correctly classified as true positive reduces the actual false positive rate to 65 in 8,000 classifications (0.8125%).

There were 88 total instances of false negatives consisting of multiple occurences of 22 unique samples. While 53 of those were genuine cases of false positives, there were six misannotated unique instances accounting for 35 cases where the 'negative' classification conflicting with the gold labels should be considered appropriate.

- (11384) CRVO in a 27 y.o. is a very unusual occurrence. Did your doctors figure out what caused it?

The complete list of samples incorrectly annotated as 'positive' surfacing in our evaluation series is *14485, 11520, 11522, 11194, 11960*.

Given the significant share of debatable gold labels in both false positive and false negative classifications, we suggest revising the dataset based on an investigation of the aforementioned error patterns of repeated evaluation runs.

5 Conclusion

We applied state-of-the art transformer language models to three different shared tasks, succesfully employing dataset fusion to broaden the training base for two closely related tasks. We observed on par results between models in the vicinity of human performance on the gold annotation. The resulting model for task 4 achieved F1 scores of 0.912 (P 0.924, R 0.901) and 0.917 (P 0.891, R 0.904) on the blind test set, outperforming both mean and median F1 scores (0.847, 0.869) of all submissions in the task by a significant margin. These results, despite generalization loss, are also on par with the results of the datataset authors' benchmark F1 on the validation set of 0.914 (Klein et al., 2022). Our best model for task 9 based on the same strategy scored 0.956 F1 on the blind test set (P=0.948, R=0.963), making it the best performing contribution in the competition.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019.

FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

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. 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.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Ari Z Klein, Arjun Magge, and Graciela Gonzalez-Hernandez. 2022. Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets. *PloS one*, 17(1):e0262087.

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. Cite arxiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching1. 18(7-8):581–618.

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 –.

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.