DILAB at #SMM4H 2024: Analyzing Social Anxiety Effects through Context-Aware Transfer Learning on Reddit Data

Sheikh Ayatur Rahman
BRAC University
Dhaka, Bangladesh
sheikh.ayatur.rahman@g.bracu.ac.bd

Azmine Toushik Wasi
Shahjalal University of Science and Technology
Sylhet, Bangladesh
azmine32@student.sust.edu

Abstract

This paper illustrates the system we design for Task 3 of the 9th Social Media Mining for Health (SMM4H 2024) shared tasks. The task presents posts made on the Reddit social media platform, specifically the *r/SocialAnxiety* subreddit, along with one or more outdoor activities as pre-determined keywords for each post. The task then requires each post to be categorized as either one of *positive*, *negative*, *no effect*, or *not outdoor activity* based on what effect the keyword(s) have on social anxiety. Our approach focuses on fine-tuning pre-trained language models to classify the posts. Additionally, we use fuzzy string matching to select only the text around the given keywords so that the model only has to focus on the contextual sentiment associated with the keywords. Using this system, our peak score is 0.65 macro-F1 on the validation set and 0.654 on test set.

1 Introduction

Analyzing health-related topics from social media data, such as Twitter, Facebook, and Reddit, to gauge public sentiment is an area of significant research interest. The SMM4H 2024 (Xu et al., 2024) shared tasks encourage researchers to address some of these research problems.

We decided to take part in Task 3. The focus of Task 3 is on **S**ocial **A**nxiety **D**isorder (SAD) (Leigh and Clark, 2018), and the motivation behind it is that while a significant number of people may experience SAD in their lives, they may experience its symptoms for much longer before actually seeking professional help. However, people often turn to social media to discuss symptoms of SAD such as the *r/socialanxiety* subreddit. In particular, this task aims to understand the effects of outdoor activities on the symptoms of SAD.

Our approach to this involved stripping text from each post so that only the context surrounding the keywords was fed into our model. Our model consisted of a RoBERTa backbone, followed a dense

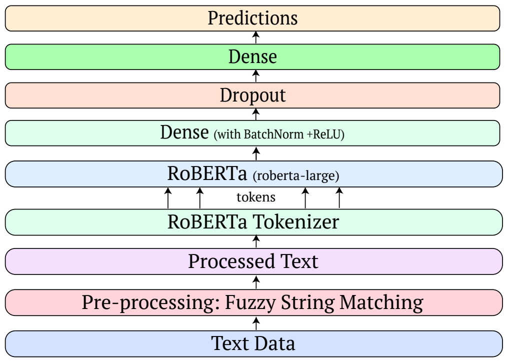

Figure 1: Model architecture

layer, a batch normalization layer, a dropout layer, and finally a classification layer, which will classify the post into one of four categories depending on the effect of the outdoor activity on symptoms of SAD : *positive*, *negative*, *no effect*, or *keyword is not an outdoor activity*. We then proceeded to fine-tune this model on that data provided.

2 System Description

Dataset. The data provided consisted of posts made on the *r/socialanxiety* subreddit, along with one or more outdoor activity keywords, and a class label for each post. In total, there were 1,800 posts in the training set, 600 posts in the validation set, and 600 posts in the test set.

Keyword	Text	Class
run	21/m. I want to experience young love, but I've never had a relationship before... (continued)	0

Table 1: Example of a data point

Pre-processing with Fuzzy String Matching. Using the *fuzzywuzzy* library, we select all instances of each keyword that appeared in the posts using Fuzzy String Matching (FSM) using Levenshtein Distance (Miller et al., 2009). To limit false positives, we only select matched keywords whose length was at least 3, and whose *similarity_score* returned from the *fuzzywuzzy.process.extract* func-

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 13–16
August 15, 2024

Model	FSM	Dropout	Tr Ac.	Tr F1	Tr Pr.	Tr Rc.	Val Ac.	Val F1	Val Pr.	Val Rc.
RoBERTa-base	N	0.3	**0.993**	**0.992**	**0.992**	**0.991**	0.728	0.569	0.585	0.556
RoBERTa-base	N	0.4	0.986	0.980	0.979	0.980	0.728	0.577	0.610	0.563
RoBERTa-base	Y	0.3	0.981	0.969	0.974	0.965	0.773	**0.653**	0.644	0.665
RoBERTa-base[t]	Y	0.4	0.977	0.966	0.965	0.967	0.762	0.637	0.620	0.662
RoBERTa-Large[t]	N	0.3	0.966	0.940	0.948	0.934	0.750	0.624	0.6200	0.632
RoBERTa-Large	N	0.4	0.977	0.967	0.971	0.977	0.747	0.598	0.625	0.581
RoBERTa-Large	Y	0.3	0.961	0.963	0.961	0.969	0.750	0.641	0.619	**0.687**
RoBERTa-Large[t]	Y	0.4	0.983	0.968	0.965	0.972	**0.777**	0.651	**0.656**	0.656

Table 2: Evaluation results by our models on the training and validation set on different setups.
[t]Test submissions, Tr=Train, Val=Validation,Ac.=Accuracy, Pr.=Precision, Rc.=Recall, Y=Yes, N=No

Submission	F1	Prec.	Rec.	Acc.
RoBERTa-base	0.590	0.587	0.620	0.633
RoBERTa-L	0.631	0.617	0.657	0.670
RoBERTa-L-FSM	**0.654**	**0.654**	**0.661**	**0.693**
Task Mean	0.5186	0.5649	0.5379	0.5746
Task Median	0.5795	0.6300	0.5885	0.6270

Table 3: Test set performance

tion is at least 90. For each instance of a matched keyword, we only select the sentence containing the keyword, the sentence preceding it, and the sentence following it. This is used as a means to ensure the model focused only on the outdoor activity and the contextual sentiment associated with it in order to perform a classification. We also keywords at the beginning of each post to give them more impact on the model's final classification.

Model. Our model relies on a RoBERTa[1] backbone (Liu et al., 2019), from HuggingFace transformers library (Wolf et al., 2020). Each sentence is first passed into the RoBERTa backbone. Since, we treat the task as a simple sequence classification task, we perform mean-pooling on the 768-dimensional embeddings (1024 if RoBERTa-large) generated by the RoBERTa model. Then, we pass the pooled embedding into 768-dimensional (1024 if RoBERTa-large) dense layer followed by a batch normalization and ReLU layer. We then pass the output through a dropout layer before passing it through a dense layer, which classifies it as one of the four categories.

Implementation Details. Our tokenizer employs a maximum token length of 256 tokens with lowercase text processing. We fine-tune models for 20 epochs using a batch size of 8, a learning rate of $1e^{-5}$ with the Adam optimizer (Kingma and Ba, 2017) and dropout p=0.4, and the cross entropy loss function. All trials use a fixed random seed of 42, and no additional data is used.

[1]https://huggingface.co/FacebookAI/roberta-base

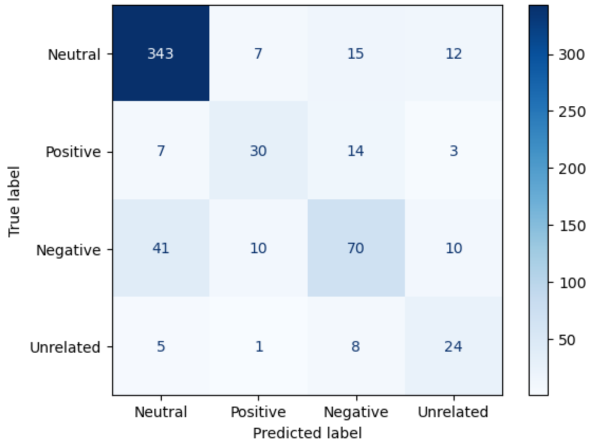

Figure 2: Raw confusion matrix of the output of RoBERTa-large with FSM on validation data

3 Results and Discussion

3.1 Training and Validation Results.

Table 2 presents the performance of different models on both training and validation data. We trained two models, RoBERTa-base and RoBERTa-large. Among these, RoBERTa-base without FSM but with dropout probability $p = 0.3$ performed best during training, achieving higher scores across all metrics. However, it showed signs of overfitting as it did not generalize well to the validation set. For RoBERTa-base, dropout probability $p = 0.3$ yielded better average scores. Conversely, for RoBERTa-large, $p = 0.4$ led to better average scores. In terms of validation data, the top-performing model was RoBERTa-large with FSM and dropout probability $p = 0.4$, achieving 77.7% accuracy and 65.6% precision. Meanwhile, the model with the highest F1 score was RoBERTa-base with FSM and dropout probability $p = 0.3$, achieving a macro F1 score of 0.653.

3.2 Impact of Text Pre-processing with Fuzzy String Matching (FSM).

In Table 2, we observe a consistent trend: models without text pre-processing using Fuzzy String

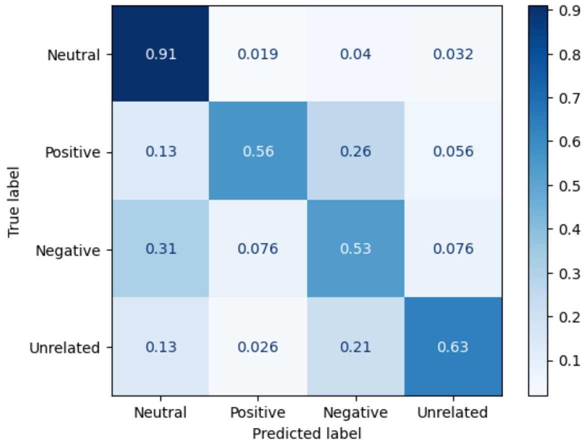

Figure 3: Normalized confusion matrix of the output of RoBERTa-large with FSM on validation data

Matching (FSM) perform better on the training data across all metrics but exhibit lower scores on the validation data. This suggests that without FSM, the models tend to overfit on noisy, unrelated data. However, with FSM, the models can focus on essential context to improve prediction accuracy.

3.3 Submissions and Test Results.

In Table 3, the test results are summarized. Among our three submissions RoBERTa-base with FSM, RoBERTa-Large with and without FSM—RoBERTa-Large with Fuzzy String Matching (FSM) performed the best, achieving an F1 score of 0.654 and 69.3% accuracy. All our submissions surpassed the mean and median task scores across all metrics.

3.4 Error analysis of our best performing model

We performed error analysis on the output of our best performing model - RoBERTa large with FSM trained using a dropout of 0.3 - on the validation set. The raw and normalized confusion matrices of the model's outputs are given in 2 and 3 respectively. We can observe that the *neutral* class (the outdoor activity has no effect), is the class on which the model performs the best. This can be explained by the imbalanced nature of the dataset. In the training set, of the 1800 training examples, 1131 belong to the *neutral* class, 160 to the *positive* class, 395 to the *negative* class, and 114 to the *unrelated* class (the keyword is not intended as an outdoor activity in the text). Due to the low numbers in the other classes, the scope for more granular analysis is small. However, a possible source inconsistency in the dataset may arise due to the fact that for examples, the correct label may be subjective. For instance consider the text - *Be a night shift stocker at Walmart. You don't really have to run the register and interact* - for which the keyword mentioned was *run*. The original label for this example was *neutral*. However, it can argued that since "run" in this scenario is not referring to the outdoor activity of running as in one would in a field or park, it should be labelled as *unrelated*. However, in this instance, "running a register" was likely considered an outdoor social activity. Subjectivity in assigning a label can lead to inconsistencies even if one person is labelling, as being consistent across multiple examples can be challenging. As such, this may lead to noisy labels and affect the model's ability to recognize patterns.

4 Conclusion

In this work, we study Social Anxiety Disorder using Reddit data to identify individuals who may experience its symptoms for a prolonged period before seeking professional help. By applying fuzzy string matching to find contexts and transfer learning with RoBERTa, we achieve a validation set macro-F1 score of 0.65 and a test set score of 0.654, outperforming the task mean and median scores. We also have analyzed errors in our best performing model with possible explanations and reasoning.

5 Limitations

While our method outperforms the mean macro-F1 score for the competition, there are some limitations to our approaches which can be addressed in future works. Some of them are listed below:

- Since we are only selecting sentences containing the keyword along with the ones preceding and succeeding them, context relevant to the keyword that are far away from the keyword is being lost. This may result in information relevant to the keyword being lost. It may be possible to design more intelligent ways of restricting context.
- We employed a very simple method to make the model focus more on the keywords - by prepending the text a few times with the keyword. It may be worthwhile to explore other ways of integrating keyword information.

References

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Eleanor Leigh and David M Clark. 2018. Understanding social anxiety disorder in adolescents and improving treatment outcomes: Applying the cognitive model of clark and wells (1995). *Clin. Child Fam. Psychol. Rev.*, 21(3):388–414.

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.

Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information theory, Computer science, String (computer science), String metric, Damerau?Levenshtein distance, Spell checker, Hamming distance*. Alpha Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Sai Tharuni Samineni, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of the 9th Social Media Mining for Health Applications Workshop and Shared Task*, Bangkok, Thailand. Association for Computational Linguistics.