CIC NLP at SMM4H 2022: a BERT-based approach for classification of social media forum posts

Atnafu Lambebo Tonja[*1], **Olumide Ebenezer Ojo**[*2],**Muhammad Arif**[*3], **Abdul Gafar Manuel Meque**[*4],**Olga Kolesnikova**[*5], **Grigori Sidorov**[*6], **Alexander Gelbukh**[*7]

[*]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico

[1]atnafu.lambebo@wsu.edu.et,{[5]kolesolga,[3] olumideoea}@gmail.com

{[4]mariff2021,[3]gafar_meque,[4]sidorov,[7]gelbukh}@cic.ipn.mx

Abstract

This paper describes our submissions for the Social Media Mining for Health (SMM4H) 2022 shared tasks. We participated in 2 tasks: a) Task 4: Classification of Tweets self-reporting exact age and b) Task 9: Classification of Reddit posts self-reporting exact age. We evaluated the two(BERT and RoBERTa) transformer based models for both tasks. For Task 4 RoBERTa-Large achieved an F1 score of 0.846 on the test set and BERT-Large achieved an F1 score of 0.865 on the test set for Task 9.

1 Introduction

Social media platforms have become more integrated in this digital era, and have impacted various people's perceptions of networking and socializing. The Social Media Mining for Health Applications (SMM4H) Shared Task involves natural language processing (NLP) challenges of using social media data for health research, including informal, colloquial expressions and misspellings of clinical concepts, noise, data sparsity, ambiguity, and multilingual posts (Gasco et al., 2022). As computational analysis opens up new opportunities for researching complex topics using social media data, models are being developed to automatically detect demographic information such as users' age (Klein et al., 2021; Tonja et al., 2022), language (Sarkar et al., 2016) (Aroyehun and Gelbukh, 2020), gender (Markov et al., 2017) (Gómez-Adorno et al., 2019), medical history (Lee et al., 2021), and so on.

More people are using social media in various ways to interact with others, share information, and express their own thoughts. Social media platforms like Twitter and Reddit have cutting-edge technology and are rich with raw, unprocessed data that can be analyzed and transformed into meaningful information. Social media research in different fields, including health (Aroyehun and Gelbukh, 2019), politics (McKeon and Gitomer, 2019), economics (Ojo et al., 2021), have included demographic variables. We participated in the social media mining Task 4 of the SMM4H 2022 shared task which centered on automatically identifying tweets that self-report the user's exact age from those that do not. In Task 9 of the same challenge, we also attempted to automatically classify Reddit posts that self-report the actual age of the online user at the time of posting from those that do not. A detailed overview of the shared tasks in the 7th edition of the workshop can be found in (Weissenbacher et al., 2022).

We applied two (BERT-Large and RoBERTa-Large) transformer models using Hugging Face [1] library. The paper is organized as follows: section 2 describes Task 4 objective, system description, experiment and result. Section 3 describes Task 9 objective, system description, experiment and result. Finally, section 4 concludes the paper and sheds some light on possible future work.

2 Task 4: Classification of Tweets Self-Reporting Exact Age

The objective of this task is automatically distinguish tweets that self-report the user's exact age from those that do not.

2.1 Data description

For Task 4: SMM4H organizers provided us with a dataset which include the Tweet ID, the text of the Tweet Object, and the annotated binary class containing 0 and 1, Tweets were annotated as "1" if the user's exact age could be determined from the tweet or annotated as "0" if the user's exact age could not be determined from the tweet. The training set consists of 8,800 tweets with 5,966 examples labeled as "0" and 2,834 examples labeled as "1". The validation dataset has 2,200 tweets with

[1]https://huggingface.co/

Proceedings of the 29th International Conference on Computational Linguistic, pages 58–61
October 12–17, 2022.

1,491 examples labeled as "0" and 709 examples labeled as "1". Thus, the dataset has a huge class imbalance, 67.7% of the texts were labeled as "0" and 32.3% were labelled as "1". To solve the class imbalance problem in the given dataset we applied the random oversampling method called the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). Before the data record is entered into the text classification model, we cleaned up and pre-processed tweets. We performed the following pre-processing steps to remove unnecessary data from the dataset, this includes removal of urls, removal of tweeter usernames and stop word removal.

2.2 System Description and Experiment

We used BERT-Large-uncased (Devlin et al., 2018) and RoBERTa-large (Liu et al., 2019) implemented using Huggingface toolkit (Wolf et al., 2020) to extract the exact age from social media posts. After pre-processing the textual data as described in section 2.1, the text sequence was tokenized using the subword tokenizer by using both BERT-large and RoBERTa-large modes with maximum text length of 180. To optimize the model, we used a relu optimizer with a batch size of 64 and a learning rate of 0.0001. We used the maximum number of epochs of 10 with early stopping based on the performance of the validation set. We also used dropout of 0.1 to regularize the model. To run our experiment we used Google colab pro + with Python programming language.

2.3 Result

We evaluated the performance of our models on the validation set and the test set, we compared our validation set performance result to check the model performance before evaluating the model in the test set for submission. We evaluated the result of the models with the F1-score for the "positive" class (i.e., tweets that self-report the user's exact age).

The results on the validation set for BERT-Large and RoBERTa-Large models are reported in Table 1. As shown in Table 1, the best performing model on the validation set for Task 4 is RoBERTa-Large model. When we observed the model performance on validation set, both the models showed less result in precision, recall and F1-score for class '1'(positive class) than class '0'. When comparing the performance of the models in test set, RoBERTa-Large model was able to achieve 0.846 F1-score on the test set as seen in Table 2.

Model	Class	P	R	F1
BERT-Large	0	0.67	0.81	0.73
	1	0.27	0.14	0.18
Accuracy				0.60
RoBERTa-Large	0	0.96	0.79	0.87
	1	0.68	0.94	0.79
Accuracy				**0.84**

Table 1: Performance of our models on Task 4 validation set (unofficial results)

Model	P	R	F1-score
RoBERTa-Large	0.804	0.891	**0.846**
BERT-Large	0.737	0.886	0.805

Table 2: Performance of our models in Task 4 in test set (official results)

3 Task 9: Classification of Reddit Posts Self-Reporting Exact Age

The objective of this task is similar with the objective described in section 2, the only difference is that this task used dataset from Reddit social media.

3.1 Data description

The datasets for this task were collected from Reddit posts, the labels were annotated in similar manner as described in subsection 2.1. The dataset is disease-specific and consists of posts collected via a series of keywords associated with dry eye disease. The training set consists of 9000 posts with 6,079 examples labeled as "0" and 2,921 examples labeled as "1". The validation dataset has 1000 posts with 686 examples labeled as "0" and 314 examples labeled as "1". As described in subsection 2.1 this dataset also has class imbalance, we used the same methods to solve the class imbalance issue as described in subsection 2.1. We also followed the same procedure for data pre-processing as described in subsection 2.1

3.2 System Description and Experiment

We used the same system description and experimental setup as in Task 4 described in section 2.2 because the objective of both tasks are the same while the difference is only in the dataset source.

3.3 Result

Similarly we evaluated the performance of selected models on Task 9 validation set before evaluating and submitting the prediction file on test set. As used in section 2.2 for Task 9 we evaluated the performance of the models with F1-score for the positive class (i.e. posts annotated as "1").

The results on the validation set for BERT-Large and RoBERTa-Large models are reported in Table 3. As shown in Table 3 the best performing model on the validation set for Task 9 is RoBERTa-Large model. RoBERTa-Large model showed better result on predicting positive class(1) than BERT-Large. When evaluating the performance of the models on the test set, BERT-Large model was able to achieve an F1-score 0.865 on the test set as seen in Table 4.

Model	Class	P	R	F1
BERT-Large	0	0.97	0.88	0.92
	1	0.78	0.94	0.85
Accuracy				0.90
RoBERTa-Large	0	0.95	0.91	0.93
	1	0.82	0.90	0.86
Accuracy				**0.91**

Table 3: Performance of our models on Task 9 validation set (unofficial results)

Model	Precision	Recall	F1-score
BERT-Large	0.797	0.946	**0.865**

Table 4: Performance of our models on Task 9 test set (official results)

4 Conclusion

In this work, we describe our team submission for Social Media Mining for Health Applications shared task 2022. We have explored an application of RoBERTa and BERT language models to the task of classification of tweets self-reporting exact age and classification of Reddit posts self-reporting exact age. Our experiments have shown that RoBERTa-Large outperforms BERT-Large in classification of tweets self-reporting exact age (Task 4) in both validation and test set. For classification of Reddit posts self-reporting exact age (Task 9), we found that RoBERTa-Large outperforms BERT-Large in validation set.

In the future, we will explore the effect of class imbalance on the performance of classification models, apply different methods to solve class imbalance and their effect on model performance.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2019. Detection of adverse drug reaction in tweets using a combination of heterogeneous word embeddings. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 133–135, Florence, Italy. Association for Computational Linguistics.

Segun Taofeek Aroyehun and Alexander Gelbukh. 2020. Nlp-cic at hasoc 2020: Multilingual offensive language detection using all-in-one model. In *FIRE*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*.

Helena Gómez-Adorno, Roddy Fuentes-Alba, Ilia Markov, Grigori Sidorov, and Alexander Gelbukh. 2019. A convolutional neural network approach for gender and language variety identification. *Journal of Intelligent and Fuzzy Systems*, 36(5):4845–4855.

Ari Z. Klein, Arjun Magge, and Graciela Gonzalez-Hernandez. 2021. Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets. *CoRR*, abs/2103.06357.

Lung-Hao Lee, Man-Chen Hung, Chien-Huan Lu, Chang-Hao Chen, Po-Lei Lee, and Kuo-Kai Shyu. 2021. Classification of tweets self-reporting adverse pregnancy outcomes and potential covid-19 cases using roberta transformers. In *SMM4H*.

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.

Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, and Alexander Gelbukh. 2017. The winning approach to cross-genre gender identification in russian at rusprofiling 2017. In *FIRE*.

Robin Tamarelli McKeon and Drew H. Gitomer. 2019. Social media, political mobilization, and high-stakes testing. *Frontiers in Education*, 4.

O. E. Ojo, A. Gelbukh, H. Calvo, and O. O. Adebanji. 2021. Performance study of n-grams in the analysis of sentiments. *Journal of the Nigerian Society of Physical Sciences*, 3(4):477–483.

Sandip Sarkar, Saurav Saha, Jereemi Bentham, Partha Pakray, Dipankar Das, and Alexander Gelbukh. 2016. Nlp-nitmz@dpil-fire2016: Language independent paraphrases detection. In *FIRE*.

Atnafu Lambebo Tonja, Muhammad Arif, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2022. Detection of aggressive and violent incidents from social media in spanish using pretrained language model.

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 *In Proceedings of the Seventh Social Media Mining for Health (SMM4H) Workshop and Shared Task*.

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.