MaNLP@SMM4H'22: BERT for Classification of Twitter Posts

Keshav Kapur
Manipal Institute of Technology
keshav29kapur@gmail.com

Rajitha Harikrishnan
Manipal Institute of Technology
rajithasuja@gmail.com

Sanjay Singh
Manipal Institute of Technology
sanjay.singh@manipal.edu

Abstract

The reported work is our straightforward approach for the shared task "Classification of tweets self-reporting age" organized by the "Social Media Mining for Health Applications (SMM4H)" workshop. This literature describes the approach that was used to build a binary classification system, that classifies the tweets related to birthday posts into two classes namely, exact age(positive class) and non-exact age(negative class). We made two submissions with variations in the preprocessing of text which yielded F1 scores of 0.80 and 0.81 when evaluated by the organizers.

1 Introduction

Determining the exact age of an individual is crucial to increasing the use of social media data for research purposes. In this contemporary world, adolescents use social media to the extent that it can have some very severe effects on their overall well-being if not monitored. Hence, a few applications like Twitter have set up some age restrictions for the well-being of an individual. They automatically detect the age of an individual trying to protect them from viewing unnecessary and harmful content.

In this work, we determine the exact age of an individual based on their tweets on Twitter. This helps in validating if a particular user has faked their age or not. One of the major challenges we faced while working with the data is that some of them could tweet about their friend's or relative's birthday which was getting misclassified. We have used BERT model which helps in the binary classification of our data. While developing our system for this task, we have discovered BERT that outperforms traditional training models.

2 Methodology

2.1 Pre-Processing

Initially, the organisers provided us with 8,800 training data and 2,200 validation data. This dataset consisted of three fields: tweet id, text of the Tweet Object and annotated binary class label(exact age present/absent). The training and validation data were later combined and it was pre-processed for further development. For preprocessing, we removed URLs, emoticons, hashtags, and mentions using a python package *tweet-preprocessor*. After that we removed: contractions from the tweets, special characters and extra spaces. Then we used python package called *Natural Language Toolkit* for removing the stop words. After these steps, we have further divided the pre-processing into two techniques: pronouns and removing pronouns.

2.2 Model

In our model, we have used BERT (*bert-base-uncased*) from the Hugging Face library as a classifier and Softmax as the activation function. In the BERT model, there is an important special token [CLS] which is used as an input for our choice of classifier. We have used the Adam optimizer to fine-tune our BERT model. We trained the model with 4 to 10 epochs which converged after 10 epochs. Learning rate of the optimizer is given by 5e - 5.The batch size used is 32.

3 Evaluation

In the validation phase, our model produced satisfactory results with about 90%. In the test data, 10,000 tweets were provided by the organizers. We have first pre-processed with pronouns and then removed pronouns in the next round of pre-processing. After evaluation, our models generated an F1-score of 0.80 and 0.81.

Table 1 shows our evaluation scores for Precision,

Proceedings of the 29th International Conference on Computational Linguistic, pages 42–43
October 12–17, 2022.

Model	Precision	Recall	F1-Score
Model 1	0.839	0.780	0.808
Model 2	0.771	0.870	0.818

Table 1: Evaluation scores

Recall, and F1-Score as provided by the organizers. Model 1 shows scores of pre-processing with pronouns and Model 2 shows scores of pre-processing with pronouns removed.

4 Conclusion

We discussed our approach to fine-tuning our BERT model on Task 4 of the 2022 Social Media Mining for Health applications shared task. As we observe from the results, the given training data was inadequate to train on a BERT model. There was an imbalance in the number of positives and negatives given in our dataset(refer to Figure 1). An interesting observation drawn from this work is that BERT models rely on huge and balanced datasets for learning patterns. Future work might consider collecting more data points for training, fine-tuning our BERT model, and applying other state-of-the-art methods like RoBERTa.

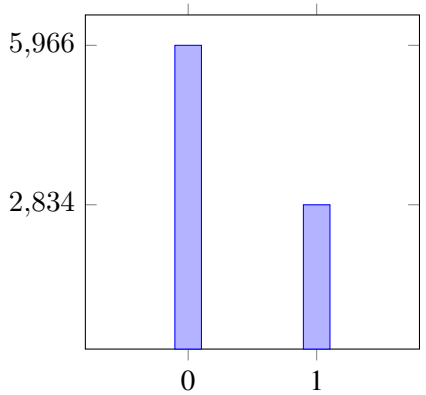

Figure 1: Summary of Dataset Labels

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

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.

Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre-Maduell, Salvador Lima Lopez, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan M Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez, editors. 2021. *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*. Association for Computational Linguistics, Mexico City, Mexico.