IITRoorkee@SMM4H 2024: Cross-Platform Age Detection in Twitter and Reddit Using Transformer-Based Model

Thadavarthi Vishnu Sri Sai Sankar, Dudekula Suraj, Mallamgari Nithin Reddy, Durga Toshniwal, and **Amit Agarwal**
Department of Computer Science and Technology
Indian Institute of Technology Roorkee
{t_vishnu,dudekula_s,mallamgari_nr,durga.toshniwal}@cs.iitr.ac.in

Abstract

This paper outlines the methodology for the automatic extraction of self-reported ages from social media posts as part of the Social Media Mining for Health (SMM4H) 2024 Workshop Shared Tasks. The focus was on Task 6: "Self-reported exact age classification with cross-platform evaluation in English." The goal was to accurately identify age-related information from user-generated content, which is crucial for applications in public health monitoring, targeted advertising, and demographic research. A number of transformer-based models were employed, including RoBERTa-Base, BERT-Base, BiLSTM, and Flan T5 Base, leveraging their advanced capabilities in natural language understanding. The training strategies included fine-tuning foundational pre-trained language models and evaluating model performance using standard metrics: F1-score, Precision, and Recall. The experimental results demonstrated that the RoBERTa-Base model significantly outperformed the other models in this classification task. The best results achieved with the RoBERTa-Base model were an F1-score of 0.878, a Precision of 0.899, and a Recall of 0.858.

1 Introduction

Social media data (e.g., Reddit and Twitter) plays a crucial role in health informatics, helping researchers understand public opinions on health-related issues. To engage researchers and students in analyzing social media data, the Social Media Mining for Health Applications (SMM4H) shared tasks workshop is organized by the University of Pennsylvania's Health Language Processing Lab.

The SMM4H-2024 workshop focuses on Large Language Models (LLMs) and the generalizability of social media NLP. This year's workshop includes seven shared tasks. Among these, we have worked on Task 6: "Self-reported exact age classification with cross-platform evaluation in English." Our motivation for tackling this task stems from the observation that many patients express their health needs and medical concerns through social media. To enhance the research utility of social media data, it is essential to develop techniques for automatically identifying demographic information, such as user age, from these platforms. A detailed overview of the shared tasks in the 9th edition of the workshop can be found in (Xu et al., 2024)

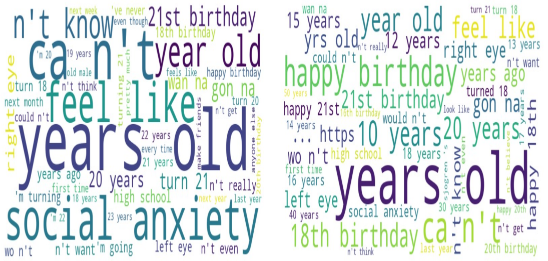
The Task-6 presented in this workshop is a continuation of the work from SMM4H 2022 (Weissenbacher et al., 2022) workshop. Several papers have already been published addressing this task, highlighting its importance and the various approaches researchers have taken to solve it. For instance, (Claeser and Kent, 2022), (Kapur et al., 2022), (Tonja et al., 2022) and (Klein et al., 2022) explored different methodologies and achieved notable results.

The structure of this paper is as follows: Section 2 describes the motivation for the transformer based approaches, Section 3 describes the dataset and details of the classification task. Section 4 outlines the methodology and various experiments conducted. Section 5 discusses the results of these experiments. Finally, Section 6 concludes the paper, summarizing our findings and suggesting potential directions for future research.

2 Motivation for Using Transformer-Based Approaches

The task of extracting self-reported ages from social media posts, specifically tweets, poses significant challenges due to the similarity in content between posts that do and do not contain self-reported age information. This is illustrated by the word clouds generated from the labeled & unlabelled dataset for the top 50 bigrams, as shown in Figures 1a and 1b.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 101–105
August 15, 2024 ©2024 Association for Computational Linguistics



(a) Word Cloud for Labeled Dataset (Self-Reported Age)

(b) Word Cloud for Unlabeled Dataset (No Self-Reported Age)

Figure 1: Word Clouds for Labeled and Unlabeled Datasets

The word clouds from the labeled (self-reported age) and unlabeled (no self-reported age) datasets appear very similar, indicating that distinguishing between the two categories based solely on content can be difficult. Common phrases such as "years old," "happy birthday," and "social anxiety" are prevalent in both word clouds, which underscores the complexity of the task.

Given the subtle differences in the language used in these tweets, transformer-based models (Kalyan et al., 2021) are well-suited for this classification task. These models leverage deep learning techniques and advanced natural language understanding capabilities to capture nuanced patterns in text. Mathematically, the problem can be formulated as follows:

Given a set of tweets $T = \{t_1, t_2, \ldots, t_n\}$, the goal is to classify each tweet t_i into one of two classes:

$$y_i = \begin{cases} 1 & \text{if the age of the user can be determined} \\ 0 & \text{otherwise} \end{cases}$$

Transformer-based models, such as RoBERTa, BERT, BiLSTM, and T5, can be fine-tuned to learn the mapping function $f : T \rightarrow Y$ where $Y = \{y_1, y_2, \ldots, y_n\}$. These models use contextual embeddings and attention mechanisms to effectively distinguish between tweets that contain self-reported age information and those that do not.

3 Task Description and Dataset

The goal of this task is to develop an effective algorithm for the binary classification of tweets based on whether the exact age of the user can be determined from the tweet at the time it was posted. A tweet is labeled as "1" if the user's age can be determined from the text of the tweet at the time it was posted; otherwise, the tweet is labeled as "0." For this task, datasets are provided by the SMM4H-2024 workshop. The dataset includes posts from two social media platforms: Twitter and Reddit. The dataset contains approximately 13,200 tweets and 100,000 Reddit posts. This Dataset is divided into three parts:

The training data consists of 8,800 tweets from SMM4H22 and 100,000 unlabeled Reddit posts. For validation, we used 2,200 tweets from SMM4H22 and 1,000 Reddit posts related to dry eye disease, also from SMM4H22. The testing data includes 2,200 tweets from SMM4H22, 2,000 Reddit posts about dry eye disease from SMM4H22, and 12,482 Reddit posts about social anxiety. The Evaluation metrics used to evaluate in this task are standard metrics F1-Score, Precision, Recall.

4 Methods and Experiments

In this task, various models from the Huggingface toolkit (Wolf et al., 2020) were used for the automatic extraction of the exact age from tweets. The models used for the experiments are listed below:

RoBERTa-Base (Liu et al., 2019): This model was trained on a dataset containing 8,800 tweets. The tweets were tokenized using the model's subword tokenizer with a maximum token length of 128. The Adam optimizer (Kingma and Ba, 2017) was used for training with the following hyperparameters: learning rate = 2e-5, number of epochs = 10, and batch size = 64.

BERT-Base (Devlin et al., 2019): The BERT model from the Hugging Face library was used as a classifier. The create optimizer module from the hugging face toolkit library was used to fine-tune the BERT model. The learning rate was set to 2e-5, and the batch size was 16. The model is trained for upto 4 epochs

BiLSTM (Graves et al., 2013): BiLSTMs are effective for binary classification tasks as they contain two separate layers, one processing the input in the forward direction and the other in the backward direction. This bidirectional processing ensures the entire text sequence is effectively classified. The Sigmoid function was used as the activation function, and the Adam optimizer was used for training. The model was trained for up to 5 epochs.

Flan T5 Base (Chung et al., 2022): This model from the Hugging Face library was used for various NLP tasks, including summarizing, question answering, and text classification. For this task, the model was fine-tuned on the training dataset with a learning rate of 3e-4, a batch size of 8, and trained for 2 epochs.

5 Results and Discussion

Using the validation dataset provided by the organizers of the SMM4H-2024 workshop, the performance of the models was evaluated using the metrics Precision, Recall, and F1-Score. The results of the different models on the validation dataset are presented in Table 1. From Table 1, it is evident that the RoBERTa-Base model achieved the highest performance with an F1-Score of 0.88, Precision of 0.899, and Recall of 0.858, demonstrating the model's robustness in accurately identifying tweets where the user's exact age can be determined.

Subsequently, the predictions of the best-performing model, the RoBERTa-Base, were submitted on the test dataset. The results of these predictions are shown in Table 2. These results further validate the effectiveness of the RoBERTa-Base model in this classification task, reaffirming its suitability for practical applications in extracting demographic information from social media posts.

Additionally, Table 3 presents the predictions of our models on some sample tweets. This table lists predictions made by BERT, FlanTS, BiLSTM, and RoBERTa models, indicating whether they correctly identified the presence of self-reported age information. The correct labels for examples 1 and 3 are "0", indicating no self-reported age, while the correct labels for examples 2 and 4 are "1", indicating the presence of self-reported age.

In this context, RoBERTa demonstrates a high level of accuracy, correctly predicting the labels for three out of the four examples. Specifically, it successfully identified examples without self-reported age information (examples 1 and 3) and one with self-reported age (example 2). However, it incorrectly classified example 4, highlighting its occasional limitations in dealing with certain nuances in the language.

Model	Precision	Recall	F1 Score
RoBERTa	0.89	0.86	0.88
BERT	0.90	0.87	0.87
BiLSTM	0.69	0.71	0.73
FlanT5	0.57	0.98	0.72

Table 1: Performance of our models on Task 6 Validation Dataset

Model	Precision	Recall	F1 Score
RoBERTa	0.899	0.858	0.878

Table 2: Performance of RoBERTa-Base model on Test Dataset

S.No	Example	Models
1	64 is for distance (use this for the glasses unless you are getting reading glasses).	BERT × FlanT5 × BiLSTM ✓ RoBERTa ✓
2	Well I'm 19 so later teens and earlier 20s sounds good to me but really I don't care, anyone can join	BERT ✓ FlanT5 ✓ BiLSTM × RoBERTa ✓
3	DMEK can give you 20/20 but not every time.	BERT ✓ FlanT5 ✓ BiLSTM × RoBERTa ✓
4	23, Indian male. Spend a lot of time in front of computer screens.	BERT × FlanTS ✓ BiLSTM ✓ RoBERTa ×

Table 3: Examples and Models' Predictions

BERT correctly predicted two examples but struggled with implicit age information. FlanTS also performed well on two examples but was less consistent than RoBERTa, especially with subtle context differences. BiLSTM, known for its sequential processing capability, accurately identified tweets without self-reported age information but had difficulty with context-dependent tweets.

Overall, Table 3 illustrates the comparative performance of these models, with RoBERTa generally showing superior accuracy but still facing challenges with certain tweet constructs. This comparative analysis underscores the importance of leveraging transformer-based models for their advanced contextual understanding capabilities.

6 Conclusion

This work presents our experiments on the binary classification of texts to determine whether they contain self-reported exact age information. We explored various transformer-based models, including RoBERTa-Base, BERT-Base, BiLSTM, and Flan T5 Base aiming to identify the most effective model for this binary classification task. Our results demonstrate that the RoBERTa-Base model outperforms the other models, achieving an F1-score of 0.878, a Precision of 0.899, and a Recall of 0.858.

The superior performance of the transformer-based models underscore their potential for practical applications in public health monitoring, targeted advertising, and demographic research. By leveraging the advanced natural language processing capabilities of transformer-based models, we were able to effectively capture nuanced patterns in text, thereby improving the accuracy of age classification tasks.

While the results are promising, there are still challenges to address, such as the occasional misclassification of tweets with subtle context differences and the token limit of transformer-based models. Our study highlights the importance of model robustness and the need for further research in this area. Future work could focus on developing effective chunking methods for input text to improve the classification accuracy and dealing with implicit age information and diverse linguistic constructs.

Overall, this study demonstrates the efficacy of transformer-based models in classifying social media posts based on the presence of self-reported age information, providing a foundation for more advanced analysis of user-generated content.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Daniel Claeser and Samantha Kent. 2022. Fraunhofer FKIE @ SMM4H 2022: System description for shared tasks 2, 4 and 9. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 103–107, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing. *Preprint*, arXiv:2108.05542.

Keshav Kapur, Rajitha Harikrishnan, and Sanjay Singh. 2022. MaNLP@SMM4H'22: BERT for classification of Twitter posts. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 42–43, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Ari Klein, Arjun Magge, and Graciela Gonzalez. 2022. Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets. *PLOS ONE*, 17: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. *Preprint*, arXiv:1907.11692.

Atnafu Lambebo Tonja, Olumide Ebenezer Ojo, Mohammed Arif Khan, Abdul Gafar Manuel Meque, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2022. CIC NLP at SMM4H 2022: a BERT-based approach for classification of social media forum posts. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 58–61, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, 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 Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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. 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.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, 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 Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.