712forTask7 at #SMM4H 2024 Task 7: Classifying Spanish Tweets Annotated by Humans versus Machines with BETO Models

Hafizh R. Yusuf,[1] David Belmonte,[2] Dalton Simancek,[3] V.G.Vinod Vydiswaran[3,1]
[1]School of Information, [2]Department of Psychiatry, [3]Department of Learning Health Sciences
University of Michigan, Ann Arbor
{hafizhry, dbelmont, daltonsi, vgvinodv}@umich.edu

Abstract

The goal of Social Media Mining for Health (#SMM4H) 2024 Task 7 was to train a machine learning model that is able to distinguish between annotations made by humans and those made by a Large Language Model (LLM). The dataset consisted of tweets originating from #SMM4H 2023 Task 3, wherein the objective was to extract COVID-19 symptoms in Latin-American Spanish tweets. Due to the lack of additional annotated tweets for classification, we reframed the task using the available tweets and their corresponding human or machine annotator labels to explore differences between the two subsets of tweets. We conducted an exploratory data analysis and trained a BERT-based classifier to identify sampling biases between the two subsets. The exploratory data analysis found no significant differences between the samples and our best classifier achieved a precision of 0.52 and a recall of 0.51, indicating near-random performance. This confirms the lack of sampling biases between the two sets of tweets and is thus a valid dataset for a task designed to assess the authorship of annotations by humans versus machines.

1 Introduction

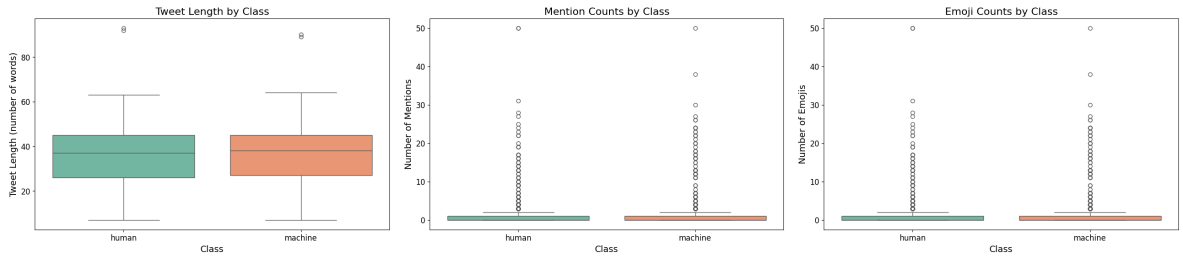
Tweeting has become a significant way for people to connect and communicate with each other individually and as a community. In 2021, 23% of adults in the United States reported to using Twitter (Dinesh and Odabaş, 2023). This type of social media is used for a variety of reasons, including news, entertainment, communication between family and friends, information on brands, and for developing a professional network. For example, COVID-19 spurred Twitter users to share their personal experiences, emotions, and beliefs during a turbulent and confusing time of a global epidemic (Cuomo et al., 2021).

Given the prevalence of its use, tweet content can provide information for understanding public sentiment and identifying specific areas of concern. However, sifting through the volumes of tweets can be an arduous task. Machine learning provides an automated approach for analyzing tweets. Leveraging its tremendous processing speed, a machine learning model, such as BERT and GPT-3, can be trained to identify specific patterns and annotate features of interest quickly compared to manual annotation by humans (Ding et al., 2023).

The #SMM4H 2024 Shared Task 7 builds over the dataset from Task 3 of the 2023 Social Media Mining for Health. Task 3 involved a dataset of 10,150 tweets describing COVID-related symptoms (Klein et al., 2023). The dataset was annotated by human experts who were medical doctors and also native speakers of Latin-American Spanish. The 2023 challenge involved training a machine learning model that could identify and extract symptoms in the tweets that were either personally described or mentioned by a third party. The leading model achieved an F1 score of 0.94 for identifying the character offsets of COVID-19 symptoms.

The dataset provided for the #SMM4H 2024 Shared Task 7 utilized some of the tweets from #SMM4H 2023 that were annotated by either human or machine (Xu et al., 2024). The intended aim was to identify whether the tweet was annotated by a human or a machine. Due to the lack of additional tweet annotation data for this classification task, we first conducted an exploratory data analysis to investigate if the two subset of data – those annotated by humans versus machines – were fundamentally different. Then, we trained a BERT-based classifier to explore differences in content between the two subsets of tweets. Our primary research goal was to confirm the absence of sampling biases that could affect the training of a classifier designed to distinguish between human and machine annotations.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 149–152



(a) (b) (c)

Figure 1: Comparative analysis of tweet characteristics by class and (a) tweet length, (b) mentions, and (c) emojis.

2 Exploratory Data Analysis

The Task 7 dataset included only the tweet text and an associated label indicating the author of the annotation, without any additional text annotations. While our initial understanding of the task was that the model should distinguish between annotations generated by humans versus machines, the absence of any generated annotations confirmed the overall purpose of the task. We framed our work to leverage the available data and examine the differences between tweets annotated by humans and those annotated by machines. We trained a binary tweet classifier to verify that there are no significant differences between the tweets in the two annotation groups. This aims to ensure that the tweets that were sampled for human annotation vs. machine annotation did not suffer from selection bias, and such bias did not affect the classifier trained to predict annotation authorship, as intended in the original Task 7.

We first approached the challenge by inspecting the provided dataset. There was a total of 4,603 tweets of which 2,288 were labelled "human" and 2,315 labeled "machine." The counts were sufficiently balanced. Then, we performed an exploratory data analysis to dive deeper into the two subset features.

As seen in Fig. 1, a comparison of the tweet lengths, number of mentions, and emojis did not show any significant differences between those labeled as humans or machines. This initial comparison suggests that both groups of tweets have similar textual features.

Word cloud analysis on the human and machine-annotated tweets were also similar, indicating that the overall words and their frequencies were similar between both groups (Fig. 2). Finally, analyzing the bi-grams, which include pairs of consecutive words, also didn't show any significant differences

Figure 2: Comparative analysis of tweet word cloud by class

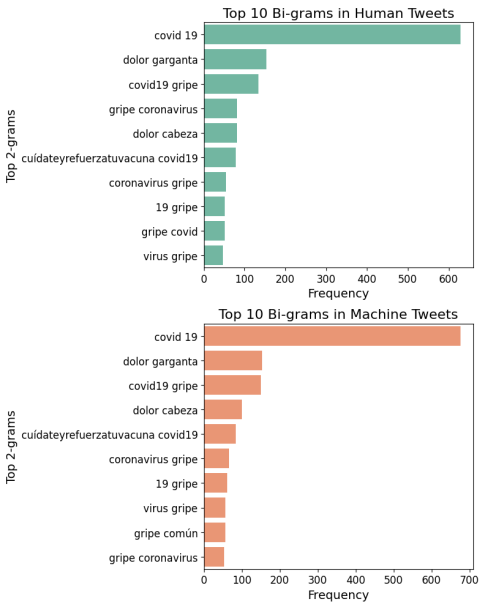


Figure 3: Comparative analysis of tweet bi-grams by class

between the two groups (Fig. 3). This further supports the hypothesis that the language content of the tweets was similar regardless of whether humans or machines annotated them.

3 BERT Model Selection and Fine-tuning

Data Pre-processing: After conducting the exploratory data analysis, we found that the number of emojis and mentions was similar between the two groups. Therefore, we removed emojis and mentions from the tweets to enable the model to focus on the core textual content. We retained the stop words to preserve the context of the tweets, which is beneficial for training a large language model.

Model training As the tweets were written in Spanish, we trained a large language model (LLM) using BETO, a BERT-based model for Spanish text (Cañete et al., 2020). This LLM was also referenced in the #SMM4H 2023 task overview (Klein et al., 2023). We fine-tuned BETO by adjusting its hyperparameters for optimal results.

For our initial submission, we found that the optimal parameters for the fine-tuned BETO model were as follows: training epochs=5, train batch size=8, evaluation batch size=32, learning rate=5e-5, and weight decay=0.01. In the following sections, this is referred to as Configuration 1.

After the evaluation period ended, we tried to further improve the model's performance and consistency by training it with various parameters and defining the seeds for the Transformers, PyTorch, and Numpy libraries. Additionally, to gain more robust metrics performance, we trained the model three times and submitted each of the runs to get their precision and recall scores. The best alternate configuration was as follows: training epochs=15, train batch size=16, evaluation batch size=32, warmup steps=100, learning rate=1e-6, and weight decay=0.01. This is referred to as Configuration 2.

4 Results

As shown in Table 1, Configuration 1 yielded a precision of 0.52 and a recall of 0.51 on the test set. Based on the scores scared by the organizers, this placed our submission at the top of the leaderboard.

In the post-evaluation phase, we found that Configuration 2 yielded better performance and consistency, with an average precision of 0.51 and a recall of 0.54, which was a slight increase compared to the initial submission. Tuning other hyperparameters did not yield any significant improvement.

Configuration	Precision	Recall	F1 score
Configuration 1	0.52	0.51	0.51
Configuration 2	0.51	0.54	0.52

Table 1: Model performance based on two parameter configurations

5 Discussion

Despite initial misunderstanding of the task, we realized that the actual aim of the shared task was to determine whether the dataset used in previously published work on generating machine-based annotations incurred any training bias. The dataset provided for training and validation did not contain any additional information aside from the tweet text and label. Particularly, there was no information about the specific COVID symptoms from the human or machine annotator that may have been informative in distinguishing between the human and machine annotations.

Based on our results, our models were unable to find any key distinguishing factors separating human-annotated and machine-annotated tweets. This strengthens the conclusions of the baseline system mentioned in the challenge, which achieved an 82% classification accuracy for a tweet annotation classifier. Our best models, formulated as tweet authorship classifiers, only achieved a random-chance performance, indicating that the tweets labeled by machines were indistinguishable from those labeled by humans. This aligns with our exploratory data analysis, which also revealed that the features of the two subsets of tweets are very similar.

6 Conclusion

Our best fine-tuned BETO model achieved an overall F1 score of 0.52, and was slightly better than our submitted run, which achieved an F1 score of 0.51. Our submitted run achieved the highest performance on #SMM4H 2024 Task 7 to distinguish between tweets that were annotated by humans vs. machines. Our findings from exploratory data analysis and training a classifier between the two subsets of tweets indicate no sampling biases and help confirm the applicability of this dataset in training a tweet annotation authorship classifier.

Limitations

Our study faced several limitations that should be noted. Firstly, the dataset was exclusively in Latin-American Spanish, which limited the author's understanding because none of them were native Spanish speakers. Additionally, we encountered the absence of detailed tweet annotations, particularly specific COVID-19 symptoms identified by human or machine annotators, further limiting our ability to distinguish between human and machine annotations effectively. These constraints may have influenced our model's performance and its alignment with the intended research objective of the #SMM4H shared task. Future research should incorporate detailed tweet annotations to enhance task clarity and model effectiveness.

Ethics Statement

In the Task 7 dataset of the #SMM4H 2024, which includes tweets referencing COVID-19 symptoms, the authors recognize these tweets as private data that is publicly accessible under the ongoing consent provided by Twitter/X's User Agreement and Terms and Conditions. The tweets for this study were sourced through the #SMM4H 2024 shared task coordinators and were accessed under the guidance of an academic advisor strictly for research purposes. The data was exclusively utilized for participation in #SMM4H 2024 Task 7 and for no other uses.

References

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *Practical ML for Developing Countries Workshop @ ICLR 2020*, pages 1–9.

Raphael E. Cuomo, Vidya Purushothaman, Jiawei Li, Mingxiang Cai, and Tim K. Mackey. 2021. A longitudinal and geospatial analysis of COVID-19 tweets during the early outbreak period in the United States. *BMC Public Health*, 21(1):793. DOI:10.1186/s12889-021-10827-4.

Shradha Dinesh and Meltem Odabaş. 2023. 8 facts about Americans and Twitter as it rebrands to X. *Pew Research Center*.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Ari Z. Klein, Juan M. Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Jesus Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. 2023. Overview of the 8th social media mining for health applications (#SMM4H) shared tasks at the AMIA 2023 annual symposium. *medRxiv*. Preprint. PMID: 37986776; PMCID: PMC10659479.

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.