ARGUABLY@SMM4H'22: Classification of Health Related Tweets Using Ensemble, Zero-Shot and Fine-Tuned Language Model

Prabsimran Kaur
Thapar University, Patiala, India
pkaur_be18@thapar.edu

Guneet Singh Kohli
Thapar University, Patiala, India
guneetsk99@gmail.com

Jatin Bedi
Thapar University, Patiala, India
jatin.bedi@thapar.edu

Abstract

With the increase in the use of social media, people have become more outspoken and are using platforms like Reddit, Facebook, and Twitter to express their views and share the medical challenges they are facing. This data is a valuable source of medical insight and is often used for healthcare research. This paper describes our participation in Task 1a, 2a, 2b, 3, 5, 6, 7, and 9 organized by SMM4H 2022. We have proposed two transformer-based approaches to handle the classification tasks. The first approach is fine-tuning single language models. The second approach is ensembling the results of BERT, RoBERTa, and ERNIE 2.0.

1 Introduction

A rapid increase in the use of social media has been seen in the past decade. Social media platforms such as Twitter, Reddit, and Facebook have become a place for people to articulate their views and emotions. Twitter especially has become a medium for people to share their medical lifestyle and the health-related problems that they are facing. Thus making Twitter an essential resource for extracting meaningful data that can help better understand and improve health services. The advancement of Natural Language Processing (NLP) in deep neural models and its ability to effectively process and understand data has attracted the attention of the healthcare research community. These healthcare researchers have developed a keen interest in processing this available data efficiently using deep learning.

The Social Media Mining for Health Applications (SMM4H) (Davy Weissenbacher, 2022) aims to bring researchers worldwide for the mining, representation, and analysis of data related to health, such as updates regarding COVID-19 and its vaccination status, drugs, and medical treatments that can help gain medical insights. This year SMM4H has proposed ten tasks that involve data classification, extraction, and Named Entity Recognition. Our team has participated in various classification tasks, namely, Task 1a, 2a, 2b, 3, 5, 6, 7, and 9.

Task 1 focuses on better understanding Adverse Drug Reactions (ADRs). The aim of Task 1a was to distinguish tweets mentioning adverse drug effects (ADE) from other tweets (NoADE). Task 2a focused on determining an author's stance toward various issues related to COVID-19. The training data were annotated for perspective according to three categories: favor (positive stance), against (negative stance), and neither (neutral stance). Task 2b focused on identifying whether a tweet contains a premise (a statement that can be used as an argument in a discussion), where "1" indicates that the tweet has a premise (argument), and "0" means that the tweet doesn't contain a premise. Task 3 focused on designing a binary classifier to detect Twitter users who self-declare that they are changing their treatment medications despite being advised by a health care professional to follow the prescription.

Task 5 deals with identifying personal mentions of COVID-19 symptoms that have been tweeted in Spanish. The dataset needed to be classified into non-personal_reports for non-personal reports, LitNew/s_mentions for news and literature mentions, and Self_reports for self-reports. Task 6 focuses on distinguishing the self-reported COVID-19 vaccination status, which is labeled as "Self_reports," and users discussing vaccination status in general, labeled as "Vaccine_chatter". Task 7 (Al-Garadi et al., 2022) deals with identifying victims of Intimate partner violence (IPV) who seek help on social media like Twitter. The label "1" indicates a self-reported IPV, and "0" shows an average Tweet about domestic violence. Task 9 (Schmidt et al., 2022) focuses on the detection of demographic information on social media and distinguishing the Reddit posts that self-report the exact age of the social media user at the time of posting (annotated as "1") from those that do not (annotated as "0").

Proceedings of the 29th International Conference on Computational Linguistic, pages 138–142
October 12–17, 2022.

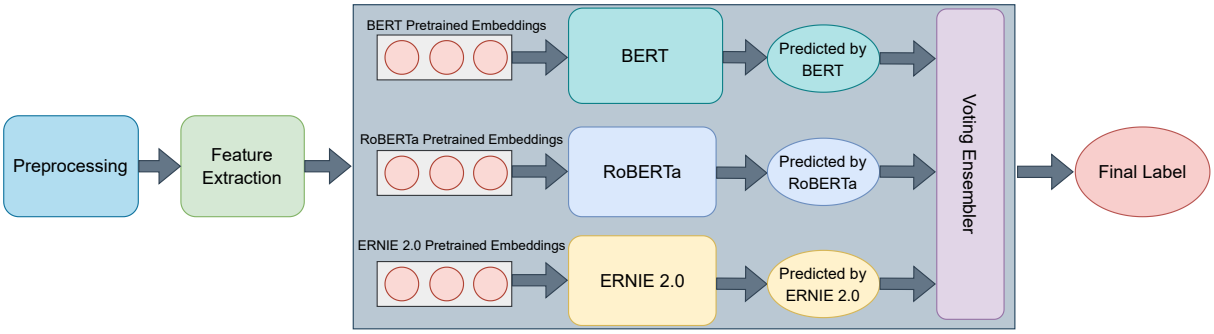

Figure 1: Architecture of Boosted Voting Ensembler

We propose two transformer-based (Vaswani et al., 2017) approaches for the classification of all the aforementioned tasks. The first approach is fine-tuning existing transformer models. The second approach uses a voting-ensemble model that comprises fine-tuned BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and ERNIE 2.0 (Sun et al., 2020). The XNLI (Conneau et al., 2018) model (zero-shot) was used to address the data imbalance in Task 1a. The multilingualism of the Spanish dataset in Task 5 was handled using XLM-RoBERTa model (Conneau et al., 2019).

2 Methodology

This section describes a detailed explanation of the approaches we have used for handling all the classification tasks. This paper proposes two architectures, all of which follow a common data preprocessing (Section 2.1).

2.1 Data Preprocessing

Social Media comments often consist of unstructured data containing special characters and emojis. Thus, a basic preprocessing involving the removal of stop words, punctuations, and emojis using the NLTK (Loper and Bird, 2002) library was performed for all the four methodologies mentioned. In case of twitter, the tweets tends to include lots of noise because of the usernames, keywords like *RT,FAV*, mentions and URLs. These redundant information were also handled using NLTK and Python pattern matching.

2.2 Fine-Tuned Transformer

The features of this preprocessed data are extracted, then each sentence is tokenized, and these tokens are mapped with their respective word IDs. The following series of steps are followed for all the sentences: a) sentence tokenization, b) prepending of [CLS] token to the start, c) appending of [SEP] token at the end, d) mapping of tokens to their word ID, e) padding or truncation of a sentence depending on the maximum sequence length, and f) mapping of the attention mask. The maximum sequence length used for each case was determined by finding the average length of the text in the dataset. The generated sequence, along with its attention mask, is then encoded. The encoded sentences are processed to yield contextually rich trained embeddings. Afterward, we pass these encodings through the desired transformer models. The transformer models used for the classification tasks were BERT, RoBERTa, ERNIE 2.0, XLM-RoBERTa, Bio Med RoBERTA (Gururangan et al., 2020), and Bio-Clinic BERT (Alsentzer et al., 2019).

2.3 Voting Ensembler

The ensemble is a learning technique in which a collection of neural networks are trained for the same task (Sollich and Krogh, 1995). The generalization ability of a neural network can be significantly enhanced by ensembling a number of neural networks (Hansen and Salamon, 1990). Ensembling involves training many neural networks and combining their predictions. The remarkable performance of this technique has made it popular in both neural network and machine learning techniques (Kohli et al., 2021).

While there are various methods of ensembling neural networks, we used the following technique. BERT, RoBERTa, and ERNIE 2.0 were individually fine-tuned (Section 2.2), after which the labels predicted by each model for each sentence were extracted. A voting system is then applied to these extracted labels, and the label which occurs the maximum number of times is selected as the final label for that sentence, as depicted in Figure 1.

Task	Technique Used	Precision	Recall	F1 Score
Task 1a	**XNLI**	**0.8395**	**0.8201**	**0.8294**
	Ensemble	0.8313	0.7775	0.8003
Task 2a	**RoBERTa**	**0.7414**	**0.7371**	**0.7384**
	Ensemble	0.7186	0.7186	0.7185
Task 2b	**BERT**	**0.7705**	**0.7684**	**0.7694**
	Ensemble	0.763433	0.7641	0.763467
Task 3	**Bio_Med**	**0.6782**	**0.5791**	**0.6028**
	Bio Bert	0.6833	0.5732	0.5967

Task	Technique Used	Precision	Recall	F1 Score
Task 5	**XLM-R**	**0.7612**	**0.7500**	**0.7534**
	XNLI	0.7387	0.7313	0.7349
Task 6	**BERT**	**0.9383**	**0.8419**	**0.8823**
	Ensemble	0.9192	0.8300	0.8723
Task 7	**RoBERTa**	**0.7639**	**0.7337**	**0.7475**
	-	-	-	-
Task 9	**RoBERTa**	**0.9307**	**0.9386**	**0.9345**
	Ensemble	0.924433	0.9327	0.928367

Table 1: Macro-Average Precision, Recall, and F1 Score to perform validation analysis with proposed methodologies

Task	Test Results	Precision	Recall	F1 Score
Task 1a	Submission	0.677	0.297	0.413
	Mean	0.646	0.497	0.562
Task 2a	Submission	-	-	0.501
	Median			0.550
Task 2b	Submission	-	-	0.6213
	Median			0.6472
Task 3	Submission	0.585	0.617	0.557
	Median	0.585	0.617	0.557

Task	Test Result	Precision	Recall	F1 Score
Task 5	Submission	0.83	0.83	0.83
	Median	0.84	0.84	0.84
Task 6	Submission	0.76	0.87	0.68
	Median	0.77	0.9	0.68
Task 7	Submission	0.784	0.689	0.734
	Median	0.790	0.716	0.763
Task 9	**Submission**	**0.896**	**0.941**	**0.918**
	Median	**0.896**	**0.019**	**0.891**

Table 2: Results released by organisers and its comparison with the mean [in Task1] and median scores of the tasks

3 Results and Discussion

3.1 Final Evaluation

Table 2 reports the final results obtained by our best systems and its comparison with the median scores from the task. Our system performed well in Task 9 where it reported a higher f1 score from the median by 0.027. For Task 3a our score was reported as median score. In Task 2a, 2b, 5, 6, 7 our submission was comparable to the arithmetic median scores that were released by the organisers. In task 1a our system is able to outperform the mean score in precision by 0.031. The closeness to the median scores in all the tasks shows that the models could have performed better with accurate hyperparameter tuning. These evaluations helped us in understanding the various shortcomings of our proposed system.

3.2 Validation Study

The hyperparameters were standardised across all of the tasks to allow for experimentation with the suggested techniques. The models were tested on total four checkpoints after being trained on two epochs with two learning rates (2x10^{-5}, 3x10^{-5}). The model that performed the best was chosen for further analysis. The top-performing fine-tuned language model and related Ensemble model for Tasks 1a, 2a, 2b, 6, and 9 have been reported on the Validation Set in Table 1. We used the BIO adjusted versions of RoBERTa and BERT for Task 3a, and only RoBERTa was reported for Task 7. To address the multilingualism challenge in Task 5, we used XLM-RoBERTa.

Table 1 helps us in understanding the validation performance of individual submission.The results reported are macro average Precision, Recall, F1 score since we wanted to give equal contribution to each class. Weighted average was avoided since the imbalance of data resulted in introduction of bias. In general it can be observed that the Ensemble models failed to outperform the Single fine tuned model in Task 1a, 2a, 2b, 6, and 9. This trend highlights the lack of robustness in Ensemble model in the given tasks.This is possible due to lack of performance of 2/3 models in the Voting Ensemble which drags down the overall result. For Task 1a and Task 5 the use of Zero Shot model was also employed to test its performance with imbalanced data.The zero shot technique outperforms in Task 1a but in general the results produced were comparable to single fine tuned language models which also helps us in realising the shortcomings of Meta Learning techniques.For task 2a,7,9 RoBERTa generates best results with F1 reaching 0.7384, 0.6028, 0.9345 respectively. For task 2b,6 BERT became the best performing model with F1 scores of 0.7694 and 0.8823 respectively. The token length was selected by calculating a 25% variation from the mean length of all the text instances available.

4 Error Analysis

This section describes the qualitative analysis of the labels predicted by the proposed architectures, as seen in Table 3. In the first instance the label predicted is noADE which indicates that the model could not understand the semantic of the sentence and focused more on the individual words like

Task	Text	Original Label	Predicted Label
Task 1	This vimpat shit is working but the side effects are hell	ADE	noADE
	i started shaking so i had to eat something :(but now i just had some tylenol pm so hopefully i'll go straight to sleep	noADE	ADE
Task 2a	(stay at home) support our P.M. and promise to follow lockdown. I request Govt to kindly postponed BANK EMJ TILL LOCKDOWN.	FAVOR	NONE
	(stay at home) Stay at home, relax back on your couch and find your dream home. Click here to explore your options.	NONE	AGAINST
Task 2b	Another GREAT perk of wearing a mask is that you can curse at people under your breath in public and they can't read your lips!!	0	1
	My daughter is immune compromised due to a rare genetic disorder, Trump can sacrifice his and the republicans children, but not my baby girl!	1	0
Task 3	Just get cortisone shot neck cool	1	0
	LisainLouKY doc gave ok take one dose excederin migraine today I took it slept woke headache	0	1
Task 5	vengo desde hace días con un dolor que puedo de la espalda alcanzó el coronavirus pero si la escoliosis	non-personal_reports	Self_reports
	El coronavirus empieza con Diarrea Quien sepa pueda decir algo Tengo un familiar con diarrea perdida del olfato gusto fiebre malestar en los huesos	non-personal_reports	Lit-News_mentions
Task 6	The vaccineinduced reduction anxiety starting creep out Good thing I good reason feel anxious cortisol levels spike	Self_reports	Vaccine_chatter
	Just got injection mean injected vaccineCovidVaccine	Vaccine_chatter	Self_reports
Task 7	TeamGivingCom My husband helped friend currently domestically violentabusive relationship We helping get situation	0	1
	johnpavlovitz Deciding testify ex trial domestic violence child abuse He got 35 years crimes BEST decision EVER	1	0
Task 9	DMEK can give you 20/20 but not every time.	0	1
	I wonder it i had it 15yrs ago just dryeye and now im in 40s so it went full blown cuza lak ot estrogen! Ahhh ughh.	1	0

Table 3: Qualitative Testing of data instance of respective shared tasks

"working". The model could not capture the true meaning of the second instance and rather made relations between shaking and tylenol, thus concluding that this was a case of side-effect.
Similarly for Task 2a the model failed to understand the discourse integration and focused on the second sentence only thus making it seem as though the Tweet was a neutral remark about postponing bank emj.In the fifth instance the model focused more on the words like "curse" and "breath", thus misleading it to believe that this statement is an argument. The model has completely failed to understand the semantic meaning of the sixth instance and is hence not labeled it as "0," indicating that there is no argument in this Tweet.

A similar trend can be seen in Task 3, 6, and 7 where the model has focused more on individual words like "dose","I","injected","violence", and "help" rather than understanding the true semantics of the sentence. The model performed relatively well in Task 9. However it failed to understand complex semantic like "now im in 40s" that indirectly indicates the age of the user. The model also puts special focus on numbers since those are commonly used for the indicating age and thus in the fifteenth instance it has made an incorrect prediction.

Our models did not perform well in tasks that comprised of medical discussions because they were not fully able to understand the underlying medical context. In case of Task 3, where Bio BERT and Bio_Med models were used the predictions were yet not very accurate perhaps due to the fact that these models were trained before the COVID-19 period and were thus not able to fully understand the terms that were used.

5 Conclusion

For the SMM4H Tasks we propose an ensemble model that leverages on pretrained representations from BERT, RoBERTa, Ernie 2.0, and a single fine tuned language model as our submission systems. Our system was able to report 91.8% F1 Score in Task 9 and 55.7% in Task 3. For other tasks our results were comparable to the arithmetic median released for all the tasks with F1 reaching 83% in Task 6 and 68% in Task 5. For Task 2a, 2b the scores achieved were 50.1% and 62.13% respectively.

References

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, 15:100217.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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 Graciela Gonzalez-Hernandez. Davy Weissenbacher, Ari Z. Klein. 2022. Overview of the seventh social media mining for health applications smm4h shared tasks at coling 2022. In *Proceedings of the Seventh Social Media Mining for Health (SMM4H) Workshop and Shared Task.*

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*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Lars Kai Hansen and Peter Salamon. 1990. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001.

Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. 2021. Arguably at comma@ icon: Detection of multilingual aggressive, gender biased, and communally charged tweets using ensemble and fine-tuned indicbert. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 46–52.

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. *arXiv preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Ana Lucía Schmidt, Raul Rodriguez-Esteban, Juergen Gottowik, and Mathias Leddin. 2022. Applications of quantitative social media listening to patient-centric drug development. *Drug Discovery Today*.

Peter Sollich and Anders Krogh. 1995. Learning with ensembles: How overfitting can be useful. *Advances in neural information processing systems*, 8.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.