AIR-JPMC@SMM4H'22: Identifying Self-Reported Spanish COVID-19 Symptom Tweets Through Multiple-Model Ensembling

Adrian Garcia Hernandez, Leung Wai Liu, Akshat Gupta, Vineeth Ravi, Saheed O. Obitayo, Xiaomo Liu, Sameena Shah
J.P. Morgan AI Research, New York, NY, USA
ag4482@columbia.edu, leungwai@wustl.edu,
{akshat.x.gupta, vineeth.ravi, saheed.o.obitayo,
xiaomo.liu, sameena.shah}@jpmchase.com

Abstract

We present our response to Task 5 of the Social Media Mining for Health Applications (SMM4H) 2022 competition. We share our approach into classifying whether a tweet in Spanish about COVID-19 symptoms pertain to themselves, others, or not at all. Using a combination of BERT based models, we were able to achieve results that were higher than the median result of the competition.

1 Introduction

The Social Media Mining for Health Applications (SMM4H) 2022 Shared Tasks competition (Weissenbacher et al., 2022) aim to encourage the use of Natural Language Processing for health research. This paper will describe our group's response to Task 5, which deals with the classification of Spanish-language tweets containing self-reported COVID-19 symptoms. The motivation for this task is the need to further develop the volume of natural language processing research on languages other than English. As Joshi et al. (2020) point out, only a small number of the world's 7000 languages are represented in Natural Language Processing technologies and related conferences. Task 5 identifies a need to increase the amount of NLP and social media research with regards to COVID-19 in all languages other than English.

Task 5 is of a three-way classification problem where a Spanish-language tweet has to be identified as one of three possible classes: **Self_reports**, **non_personal_reports**, and **Lit-News_Mentions**. What follows is a description the data-sets provided, various experiments performed, and our final submission results.

2 Dataset Information

We were presented with three data-sets consisting of labeled tweets to be used for training, an unlabeled validation set to check the final performance of our models, and an unlabeled test set to submit our final predictions. As we did not originally have the labels of the validation set during the validation stage, two approaches were applied to split the labeled training data into training and validation sets: a 80%-20% split and a 50%-50% split.

Dataset	Self-reports	Non-personal-reports	Lit-news-Mentions	Total
Training	1654	2413	5984	10051
Validation	572	859	2146	3577
Test	—	—	—	6850

Table 1: Distribution of Spanish-language tweets of each class in their respective data-sets. Since the test set is unlabeled, the distribution is unknown.

The motivation for the 80%-20% split was to feed as much data as possible when training while allowing us to measure the performance of our models. However, since the distribution of the training set skews towards more **Lit-News_mentions** tweets, the 50%-50% split approach is used to provide a better performance metric. Table 1 shows the distribution of the training, validation, and test sets.

3 Methods

3.1 Pre-processing

To explore the effects of the unbalanced class distribution in the original training set, three training sets were created on the training data from the 80%-20% split: an 'Equalized' training set in which the minority labels were randomly over-sampled to match the number of samples for the **Lit-News** label, an 'Over-Under' training set in which the minority labels were randomly over-sampled to contain 70% of the **Lit-News** label's samples while the **Lit-News** label was then randomly under-sampled to contain just 80% of its previous size, and a 'Reversed' training split set in which the positive label (self-reports) was randomly over-sampled to be the

Proceedings of the 29th International Conference on Computational Linguistic, pages 160–162
October 12–17, 2022.

size of the **Lit-News** label while the **Lit-News** label was then randomly under-sampled to be the size of the **Self_reports** label. No pre-processing was performed for the 50%-50% split training and validation data sets.

3.2 Models

To address this task, we utilized various versions of BERT (Devlin et al., 2019), as it is known to produce state-of-the-art results while being adaptable for many different applications. We used the cased and uncased versions of BERT-base-multilingual (Devlin et al., 2019) and Spanish-BERT (Cañete et al., 2020) as well as XLM-RoBERTa (Conneau et al., 2019).

3.3 Model Ensembling Post-processing

For the 80%-20% split, models were trained on the modified training data-sets for 5 iterations of 10 epochs each. The epoch with the best F1-score for the positive label (**Self_reports**) was kept per iteration. At the end of each model's training period, the best performing iteration was selected. Then, a majority-vote ensemble predictor was created by combining the predictions of each individual model and used to submit the predictions on our unlabeled test set.

A similar approach was used for the 50%-50% data-set split. However, models were trained for 5 iterations of 15 epochs each. Then, multiple model ensembling methods were tested: majority-vote of all 25 models, unweighted average and weighted averages, and a combination of separating per model and taking the top 5/10/15/20 models.

This equation was used to calculate the unweighted average: $\lfloor(\Sigma x)/25\rceil$, with x being the numeric value of the prediction, and the result rounded to nearest digit. This equation was used to calculate the weighted average: $\lfloor(\Sigma x f)/c\rceil$, with f being the F1-score of each particular model, and $c = 17.8153$ being the sum of all the F1-scores of all 25 models, with the result rounded to the nearest digit. Ultimately, the majority vote of all 25 models were used, as it produced the highest results.

Table 2 shows the baseline results for all models, while Table 3 shows the validation results after model ensembling. Ensembling was used because it has been shown to have a better performance than individual models (Jayanthi and Gupta, 2021).

Method \ Model	BERT$_{BASE}$-multilingual	BETO	XLM-RoBERTa
Original (80-20)	0.761 (0.756)	0.743 (0.737)	0.748
Over-Under (80-20)	0.751 (0.752)	0.751	—
Reversed (80-20)	0.758	—	—
Equalized (80-20)	0.747	—	—
Original (50-50)	0.750 (0.739)	0.731 (0.736)	0.740
Ensemble Results	80%-20%: **0.770** | 50%-50%: **0.755**		

Table 2: F1-Scores for the positive label (Self_reports) for various pre-processed data-sets. NOTE: Values in parenthesis refer the cased model. Not all models were tested on each of the different data-sets.

Submission \ Metric	F1-Score	Precision	Recall
80-20 Split Ensemble	0.749 (0.839)	0.649 (0.839)	0.883 (0.839)
50-50 Split Ensemble	0.753 (0.843)	0.657 (0.843)	0.881 (0.843)

Table 3: Final ensembling performance on validation data. NOTE: The top value is F1-Score for the positive label used during our testing, while the bottom value in parenthesis is the micro-average F1-Score used by CodaLab.

Classifier \ Metric	F1-Score	Precision	Recall
Baseline	0.90	0.90	0.90
Median	0.84	0.84	0.84
50-50 Split Ensemble	**0.85**	**0.85**	**0.85**
80-20 Split Ensemble	0.84	0.84	0.84

Table 4: Micro-averaged performance on test data.

4 Results

The results of our ensemble predictions for the test data compared with the Baseline and Median results are shown in Task 4. The two submissions have a different F1-scores when experimenting compared to the CodaLab submission because when experimenting, the F1-Score with respect to the **Self_report** label was used while the micro-averaged F1 Score was used in the CodaLab submission, which is akin to accuracy with a multi-class classification.

The 50%-50% split ensemble performed better than the 80%-20% split ensemble with regards to the micro-average F1-Score and vice-versa when considering the F1-Score for the positive label only. We hypothesize this is because the 80%-20% split ensemble was optimized to predict tweets belonging to the positive class. This led to an overall lowering of its accuracy, which is shown in its micro-averaged metrics. However, both the 80%-

20% split and the 50%-50% performed similar to the median score.

5 Conclusion

As demonstrated in our results section, the best performance was achieved through a majority-vote ensemble of all models. Possible opportunities for improvement would be exploring the usage of data augmentation techniques such as back-translation to address the class imbalance present in the training data-set. Moreover, we could explore different pre-processing techniques to remove possible noise from our data such as user mentions, hyperlinks or hashtags.

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 *PML4DC at ICLR 2020*.

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. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sai Muralidhar Jayanthi and Akshat Gupta. 2021. SJ_AJ@DravidianLangTech-EACL2021: Task-adaptive pre-training of multilingual BERT models for offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 307–312, Kyiv. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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 *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages –.