AIR-JPMC@SMM4H'22: Classifying Self-Reported Intimate Partner Violence in Tweets with Multiple BERT-based Models

Alec Candidato, Akshat Gupta, Xiaomo Liu, Sameena Shah
J.P. Morgan AI Research
aleccandidato@gmail.com, {akshat.x.gupta,
xiaomo.liu, sameena.shah}@jpmchase.com

Abstract

This paper presents our submission for the SMM4H 2022-Shared Task on the classification of self-reported intimate partner violence on Twitter (in English). The goal of this task was to accurately determine if the contents of a given tweet demonstrated someone reporting their own experience with intimate partner violence. The submitted system is an ensemble of five RoBERTa models each weighted by their respective F1-scores on the validation data-set. This system performed 13% better than the baseline and was the best performing system overall for this shared task.

1 Introduction

Intimate Partner Violence (IPV) refers to any form of physical or emotional abuse or aggression within a romantic relationship. Many people use social media, particularly twitter, as a means of self-reporting their experiences with IPV. Evidently, being able to parse through social media for these instances of self-reported IPV domestic violence is vital for finding victims most in need of resources and support. While scraping tweets for IPV-related content is relatively simple by keyword search, properly labeling and characterizing whether these tweets refer to self-reported IPV proves to be more challenging.

It is already known that using BERT models or transformer-based models already produces state-of-the-art results for the field of Natural Language Understanding (Devlin et al., 2019). In fact, shared tasks from previous years have already demonstrated successful results incorporating this technology (Magge et al., 2021). This paper elaborates on our submission for the task of classifying self-reported intimate partner violence on Twitter (in English). The existing research surrounding the provided data-set also demonstrates the effectiveness of RoBERTa models (Liu et al., 2019). Our final submission for this task is composed of a weighted ensemble of five RoBERTa-large models.

Split	Non-self-reported IPV	Self-reported IPV	Total
Train	4042	481	4523
Validation	480	54	534
Test	—	—	1291

Table 1: Frequency distribution for each dataset. The test dataset was unlabeled so the distribution is unknown.

2 Dataset

There were three different data-sets provided: training, validation, and test data-sets. The training and validation data-sets were labeled while the test data-set was not. All data-sets are composed entirely of tweets that are related to the topic of domestic violence. Among these tweets, 11% are identified as self-reported IPV (Al-Garadi et al., 2022). Table 1 shows the tweet distributions.

3 Method

This submission utilizes an ensembling of multiple trained RoBERTa models which each used their best epoch by means of F1-score to outperform the previous results by almost a tenth of a point. The majority ensemble takes the mode model prediction of all five RoBERTa guesses for a given tweet. The weighted ensemble takes a precise linear combination of all five RoBERTa guesses based on each model's initial performance with the validation data-set. We have two defined classes: non-self-reported (0) and self-reported (1). Let us define $P_e(y = 1 \mid x)$ as the probability that the ensemble predicts class $y = 1$ for a given tweet, x. We can represent this explicitly as:

$$P_e(y = 1 \mid x) = \frac{\sum_{i=1}^{n} a_i P_i(y = 1 \mid x)}{\sum_{i=1}^{n} a_i}$$

Proceedings of the 29th International Conference on Computational Linguistic, pages 135–137
October 12–17, 2022.

where P_i represents an individual model's probability for a given tweet, a_i represents the corresponding weight, and n is the number of models being included in the ensemble. In the case of a majority vote, $a_i = 1$. Our submission uses $a_i =$ F1-score. Consequently, we can represent the probability for class 0 as $P_e(y = 0 \mid x) = 1 - P_e(y = 1 \mid x)$. The final prediction chooses the class with the highest probability. The task submission uses this weighted ensembling method because it can easily be combined with other models. Other non-transformer models were initially developed to being included with the ensemble but not ultimately used for the final submission.

4 Preliminary Experiments

	BERT-base	RoBERTa-base	BERT-large	RoBERTa-large
Mean F1	0.718	0.749	0.710	0.779
Stdev	0.013	0.017	0.014	0.013

Table 2: Performance of BERT and RoBERTa classifiers on validation data. The F1 scores are averaged among all five models for each category. The standard deviation is also provided.

Five BERT-base (Devlin et al., 2019) and five RoBERTa-base models (Liu et al., 2019) were first trained to test the initial effectiveness of transformer classifiers. We saved the best performing epoch for each model, based on a general accuracy metric. Then, five BERT-large and five RoBERTa-large models were trained. These models saved their best performing epoch in terms of F1-score instead. The F1-score for each model was determined based on their performance with the validation dataset. The results are shown in Table 2.

The RoBERTa-large models performed significantly better than any other model we considered. Results from a majority voting ensemble and a weighted ensemble of five RoBERTa-large models were also calculated. They performed identically. Figure 1 shows the corresponding confusion matrix.

5 Results

The performances of the designed ensemble compared to the median Codalab results and the best performing model from the original research are shown on Table 3. The RoBERTa ensemble performs significantly better than all other models by all metrics. The system that the median performer

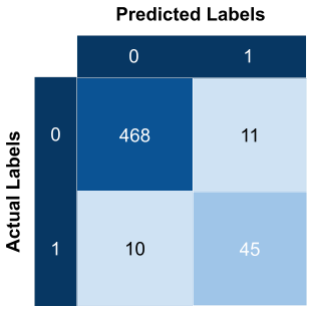

Figure 1: Confusion matrix for the RoBERTa-large ensembles on the validation data-set. This matrix represents the results for both the majority voting ensemble and weighted ensemble, since they performed identically to each other.

Classifier	F1-score	Precision	Recall
Baseline (Al-Garadi et al., 2022)	0.756	0.823	0.699
Median Task Performance	0.763	0.790	0.716
Our System	**0.851**	**0.860**	**0.841**

Table 3: Performance results for classifiers.

used is unknown at this time so no conclusions can be made comparing the results of the RoBERTa ensemble with that system. The best classifier from previous work was also built off of RoBERTa-large. Yet, this classifier achieved a 0.1 F1-score improvement. This may be because these RoBERTa models saved the epoch with the best F1-score rather than the best accuracy. Additionally, ensembling multiple transformer-based models has already proven to be effective (Jayanthi and Gupta, 2021).

6 Conclusions

This study implements ensembling with transformer-based language models to drastically improve precision and recall for this task. This overall system outperforms previous metrics by nearly 13%. This study also included initial preprocessing into part-of-speech frequency and the significance of narrative voice for self-reporting in tweets. Initial results suggests that tweets written in second or third person can help fine-tune against false positives. However, the final submission did not utilize any systems built off of this due to time limitations. Thus, these results are omitted from this paper. (Eisenberg and Finlayson, 2016) have already examined systems for narrative diegesis and point of view analysis. Ensembling transformer-based models with models trained off of narrative diegesis and point of view data may further improve results.

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*, page 100217.

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.

Joshua Eisenberg and Mark Finlayson. 2016. Automatic identification of narrative diegesis and point of view. pages 36–46.

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.

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, Salvador Lima López, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (#SMM4H) shared tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics.