CAISA@SMM4H'22: Robust Cross-Lingual Detection of Disease Mentions on Social Media with Adversarial Methods

Akbar Karimi and **Lucie Flek**
Conversational AI and Social Analytics (CAISA) Lab
Department of Mathematics and Computer Science, University of Marburg
http://caisa-lab.github.io

Abstract

We propose adversarial methods for increasing the robustness of disease mention detection on social media. Our method applies adversarial data augmentation on the input and the embedding spaces to the English BioBERT model. We evaluate our method in the SocialDisNER challenge at SMM4H'22 on an annotated dataset of disease mentions in Spanish tweets. We find that both methods outperform a heuristic vocabulary-based baseline by a large margin. Additionally, utilizing the English BioBERT model shows a strong performance and outperforms the data augmentation methods even when applied to the Spanish dataset, which has a large amount of data, while augmentation methods show a significant advantage in a low-data setting.

1 Motivation

Social media are a rich source of discussion about diseases in various contexts, including newly emerging cases, experienced symptoms, drug effects, or social consequences. Such content can be used to make more informed decisions regarding the societal perception of these diseases, as well as to augment existing medical knowledge.

Recently, pre-trained models such as BERT (Devlin et al., 2019) in the general domain and BioBERT (Lee et al., 2020) in the medical domain have become the go-to solutions for semantic annotation tasks, including previously held similar shared tasks (Klein et al., 2020; Magge et al., 2021). However, most participants have opted for a model trained on the same language (Spanish) as the task language Magge et al. (2021); Fidalgo et al. (2021); Canete et al. (2020); Ruas et al. (2021). Our research questions in this work are: (1) Can we use a domain-specific model pre-trained in one language (English) to produce reasonable performance in another? (2) Can we still further improve the robustness of such pre-trained models with methods effective on small datasets?

To address these questions, we explore the effects of adversarial and augmentation techniques, namely **adversarial training** (Goodfellow et al., 2014) and **AEDA** (Karimi et al., 2021b), on the pre-trained English BioBERT model for detecting disease mentions in Spanish tweets.

2 System Description

Our system is an ensemble of BioBERT, a bag-of-words model, improved BERT model, and adversarial data augmentation techniques:

BioBERT (Lee et al., 2020) is a BERT-based (Devlin et al., 2019) model with the same architecture trained on the general domain (Wikipedia and Book Corpus) and the medical domain (PubMed abstracts and PMC full-text articles).

BoW (Karimi et al., 2021c) or Bag-of-Words model creates a dictionary from the training words and calculates a ratio of their occurrences as a disease word. Based on this ratio (0.7 in our work), words in the test set are labeled as diseases. This is a fast and strong baseline.

Adversarial Training (AT) (Miyato et al., 2016; Karimi et al., 2021a) is a method which creates adversarial examples in the embedding space and uses them in the training process in order to make a neural network model more robust to small perturbations.

PSUM (Karimi et al., 2020) or parallel aggregation is a module that is added on top of the BERT model to improve its performance. It applies an extra BERT layer to the last four layers of the BioBERT model and aggregates the losses.

AEDA (Karimi et al., 2021b) is an adversarial data augmentation technique which inserts punctuation marks into a sentence randomly creating more data for training.

In addition to each individual method above, we also pair BioBERT, adversarial training, and PSUM models with the AEDA data augmentation technique. We acquire our final results by combining

Proceedings of the 29th International Conference on Computational Linguistic, pages 168–170
October 12–17, 2022.

labels produced for each token in the sequence using majority voting.

3 Dataset Statistics

The dataset was provided as part of the shared task of detecting disease mentions (Task 10) at the Social Media Mining for Health 2022 (SMM4H'22) (Gasco et al., 2022). The initial 5K tweets were later expanded by the organizers to make a 90K training set. The dev and test sets contained 2.5K and 23.5K tweets. Covid19 was by far the most discussed disease. The 10 most frequent disease words include *covid* (13478), *covid19* (11661), *cáncer* (7705), *covid-19* (7394), *ansiedad* (4921), *diabetes* (4253), *discapacidad* (3284), *enfermedades* (2567), *enfermedad* (2455), and *depresión* (2229).

4 Results

Table 1 summarizes our results. In order for our models to handle the labels, we convert them from the character span format into the BIO notation (with only two labels of B and O) and back.We report the performance of the models both on the BIO formatted data and the converted span match format, as this conversion process caused performance loss in our predictions due to multiple spacing, extra punctuation marks etc.

As can be seen, English BioBERT achieves the best recall. Our qualitative analysis indicates that there are already many similar words in English and Spanish in the medical entities domain. In addition, the large amount of Spanish data used for fine-tuning injects a sufficient amount of extra Spanish knowledge into the model. BioBERT also outperforms the data augmentation methods since the 90k training set seems to be large enough to neglect the augmentation benefits. However, these models seem to be more promising in low-data regimes as we can see from Table 2. While they improve the models marginally with 1000 training samples, they show significant improvements (up to nearly 10 percent) when we consider 100 tweets for training.

5 Error Analysis

A confusion matrix of the two classes indicates that the number of false positives was almost triple that of false negatives (549 vs. 192). Examining the top misclassified tokens (Table 3), we can see that the top false positives are not disease mentions, but rather prepositions that were part of the (wrongly) annotated disease spans in the training data.

	BIO			**Spans**		
Model	P	R	F1	P	R	F1
BioBERT	96.4	92.0	**94.7**	90.7	80.2	85.1
PSUM	96.8	91.4	94.0	90.8	79.9	85.0
AT	96.8	91.5	94.1	91.2	80.2	**85.4**
AEDA	96.7	91.0	93.8	90.8	79.9	85.0
PSUM+AEDA	97.0	91.2	94.0	91.1	79.9	85.1
AT+AEDA	96.9	91.1	93.9	90.8	79.5	84.8
BoW	-	-	-	76.8	57.3	65.6
Voting	-	-	-	91.4	80.1	**85.4**

Table 1: Performance on the dev set in BIO format and matching spans

	100			**1000**		
Model	P	R	F1	P	R	F1
BioBERT	61.1	73.6	66.8	86.3	89.4	87.8
PSUM	74.9	68.7	71.7	87.4	88.4	87.9
AT	62.9	72.2	67.2	86.8	89.2	**88.0**
AEDA	74.3	77.5	75.8	86.3	89.4	87.8
PSUM+AEDA	76.2	76.3	**76.2**	87.4	88.4	87.9
AT+AEDA	70.8	78.3	74.4	86.8	89.2	**88.0**

Table 2: Performance on the dev set in BIO format with 100 and 1000 training tweets

Word	de	la	#	en	y
Error frequency	41	27	25	17	13

Table 3: Top 5 errors

6 Conclusion

We introduced and successfully evaluated adversarial and augmentation strategies, which proved to be significantly influential in a low-data regime, to be combined with the BioBERT model in order to address the problem of detecting disease mentions in Spanish tweets. We also showed that the English BioBERT can have a strong performance on the same dataset; this behavior can be attributed to the dataset domain (medical entities), where the two languages share many similar words. We believe that this can be applied to other languages as well, as long as there are some similarities between them.

References

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

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.

David Carreto Fidalgo, Daniel Vila-Suero, Francisco Aranda Montes, and Ignacio Talavera Cepeda. 2021. System description for profner-smmh: Optimized finetuning of a pretrained transformer and word vectors. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 69–73.

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, and Martin Krallinger. 2022. Overview of the SocialDisNER shared task on detection of diseases mentions from healthcare related and patient generated social media content: methods, evaluation and corpora. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2020. Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021a. Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8797–8803. IEEE.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021b. Aeda: An easier data augmentation technique for text classification. *arXiv preprint arXiv:2108.13230*.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021c. Uniparma at semeval-2021 task 5: Toxic spans detection using characterbert and bag-of-words model. *arXiv preprint arXiv:2103.09645*.

Ari Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, et al. 2020. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–36.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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, et al. 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.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Pedro Ruas, Vitor Andrade, and Francisco M Couto. 2021. Lasige-biotm at profner: Bilstm-crf and contextual spanish embeddings for named entity recognition and tweet binary classification. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 108–111.