UoB at ProfNER 2021: Data Augmentation for Classification Using Machine Translation

Frances Laureano De Leon,
Harish Tayyar Madabushi and **Mark Lee**

School of Computer Science
University of Birmingham
United Kingdom

fxl846@cs.bham.ac.uk
Harish@HarishTayyarMadabushi.com, M.G.Lee@bham.ac.uk

Abstract

This paper describes the participation of the UoB-NLP team in the ProfNER-ST shared subtask 7a. The task was aimed at detecting the mention of professions in social media text. Our team experimented with two methods of improving the performance of pre-trained models: Specifically, we experimented with data augmentation through translation and the merging of multiple language inputs to meet the objective of the task. While the best performing model on the test data consisted of mBERT fine-tuned on augmented data using back-translation, the improvement is minor possibly because multi-lingual pre-trained models such as mBERT already have access to the kind of information provided through back-translation and bilingual data.

1 Introduction and Motivation

The increase of user-generated content online has allowed researchers to extract information for studies on a variety of subjects, namely tracking infectious diseases and promoting public health (Wakamiya et al., 2018; Fine et al., 2020). Consequently the emergence of COVID-19 has resulted in a rapid increase of information related to the virus on social media platforms (Zhao et al., 2020). ProfNER, a task under Social Media Mining for Health Applications (SMM4H) workshop (Magge et al., 2021), requires the identification of occupations that might be particularly affected, either mentally or physically, by the exposure to COVID-19. The task organisers give participants tweets in Spanish and English. The English tweets were translated by means of a machine translation system. Of the tweets provided, 24% contain a mention of an occupation (Miranda-Escalada et al., 2021).

Classifiers are dependent on the size and quality of the training data (Wei and Zou, 2020), and are sensitive to class imbalance. We hypothesise that increasing the number of examples in the positive class using data augmentation techniques will successfully increase the performance of trained models. This work describes the training of four classifiers using pre-trained BERT models to detect the mention of occupations in tweets. In addition to training two baseline models, BERT-Base and mBERT, we train one model on augmented textual data, mBERT-Aug, and another model on bilingual data. We compare these models to each other and to fine-tuned pre-trained BERT models, described in Section 3.2. The small increase in F1 scores over the baselines, which is inconsistent across our validation and test experiments leads us to conclude that back-translation and bilingual data input are ineffective as methods of addressing class imbalance in pre-trained models, especially multi-lingual models (See Section 4). Our models were trained using the data provided by the task organisers for subtask 7a. Results are discussed in Section 4.

2 Related Work

Augmenting textual data is challenging because it can introduce label noise and must be done before training a model (Shleifer, 2019). Among techniques developed for text augmentation is synonym replacement, random insertion, swap and deletion, as presented by Wei and Zou (2020). Shleifer (2019) uses back-translation, to translate the data in a second language and then back to the source language. They train their model on a binary classification task in a setting where low amounts of labelled data are available. Work continues to be done in back-translation for classification, as there is little research otherwise (Shleifer, 2019). In this work, we use back-translation as a tool for augmenting the text data for the positive class. This work contributes to the field of generating synthetic

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 115–117

data for text classification. Others have tried to add features to models to increase performance (Whang and Vosoughi, 2020; Lu et al., 2020), we attempt to bring together representations in different languages so as to maximise the information available to the models.

3 System Overview and Experimental Set-Up

This section describes our experimental design. The code, models, and hyper-parameters are available on our team's GitHub repository for the task [1].

3.1 Preprocessing

Punctuation, hashtags, twitter handles, emojis and URL's were all removed from the English and Spanish tweets. Tweets were tokenised using the Hugging Face Transformers library (Wolf et al., 2020).

3.2 Model Architecture

We trained four classifiers: mBERT-base, BERT-base, mBERT-Aug, and bilingual models. We utilised pre-trained mBERT-base and BERT-base to conduct our experiments (Devlin et al., 2019) using both the Spanish and English training data.

Our team fine-tuned mBERT and BERT-base to use as a baseline for our experiments. We fine-tuned both models with the 6,000 train tweets provided by the task organisers; mBERT was trained on Spanish tweets and BERT-base on English tweets. Our augmented data model is mBERT-Aug, which we trained on 6,000 Spanish tweets, and an additional 1,393 back-translated tweets. The additional tweets consist of the English data belonging to the positive class, which were translated back into Spanish using Google Translate API. We also train a bilingual model, by concatenating the output of the two transformer models. We trained this model on both the Spanish and English tweets.

4 Results and Discussion

The bilingual model obtains the best results on the validation data, while mBERT-Aug is the best scoring model on the test data, with a F-1 score of 0.83. Table 1 and Table 2 summarise the results.

We perform experiments after the evaluation period to obtain results on the test data for BERT-base and the bilingual model. We do this to compare the results of all models on the test data. We find

[1] https://github.com/francesita/ProfnerTask7a

Model	Precision	Recall	F-1
mBERT	0.8407	**0.9347**	0.89
BERT-Base	0.8763	0.8875	0.88
mBert-Aug	0.8826	0.8734	0.88
bilingual	**0.8847**	0.9194	**0.90**

Table 1: ProfNER Task 7a Validation Results

Model	Precision	Recall	F-1
mBERT	0.9538	0.7127	0.82
BERT-Base	0.6620	0.1015	0.18
mBert-Aug	0.9171	**0.7646**	**0.83**
bilingual	**0.9579**	0.6393	0.77

Table 2: ProfNER Task 7a Test Results

that neither the addition of augmented data, nor combining representations in different languages significantly improves the results, with the bilingual model performing better on the validation data, and the mBERT-Aug performing better on the test data. We believe that a reason the BERT-base and bilingual models have lower scores on the test data is due to the quality of the machine translation system that we used whereas the validation data was provided by the task organisers. For example, *Ultima Hora* in Spanish was translated as *last minute*, when it should have been translated as *breaking news* in the context it was used in. Another example is *consellera* which translates to *advisor* was not translated at all in some tweets. While these methods of data augmentation provide a small improvement, fine-tuned pre-trained BERT models are quite robust. Training on parallel corpora gave these models everything that could be extracted through back-translation and bilingual data.

5 Conclusion

Our work presents experiments with pre-trained transformer based models to perform binary classification on an imbalanced dataset. We hypothesised that the use of data augmentation and parallel inputs in multiple languages will provide a method of addressing class imbalance (Section 1). However, our experiments showed that neither of these methods are particularly powerful in this regard (Section 4). In the future, we will continue to experiment with other techniques to handle imbalanced classes, such as one-class classification and reinforcement learning-based networks to generate text.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Alex Fine, Patrick Crutchley, Jenny Blase, Joshua Carroll, and Glen Coppersmith. 2020. Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhibin Lu, Pan Du, and Jian Yun Nie. 2020. VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12035 LNCS, pages 369–382. Springer.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, 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 Applications Workshop & Shared Task*.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima López, Vicent Briva-Iglesias, Marvin Agüero-Torales, Luis Gascó-Sánchez, and Martin Krallinger. 2021. The profner shared task on automatic recognition of professions and occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Sam Shleifer. 2019. Low Resource Text Classification with ULMFit and Backtranslation. *CoRR*, abs/1903.09244.

Shoko Wakamiya, Yukiko Kawai, and Eiji Aramaki. 2018. Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study. *JMIR Public Health and Surveillance*, 4(3).

Jason Wei and Kai Zou. 2020. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Dylan Whang and Soroush Vosoughi. 2020. Dartmouth CS at WNUT-2020 Task 2: Informative COVID-19 Tweet Classification Using BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 480–484, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yi Zhao, Haixu Xi, and Chengzhi Zhang. 2020. Exploring Occupation Differences in Reactions to COVID-19 Pandemic on Twitter. *Data and Information Management*, 0(0).