Sentence Contextual Encoder with BERT and BiLSTM for Automatic Classification with Imbalanced Medication Tweets

Olanrewaju Tahir Aduragba, Jialin Yu, Gautham Senthilnathan and **Alexandra Cristea**
Department of Computer Science
Durham University, Durham, UK
{olanrewaju.m.aduragba, jialin.yu}@durham.ac.uk
{gautham.senthilnathan, alexandra.cristea}@durham.ac.uk

Abstract

This paper details the system description and approach used by our team for the SMM4H 2020 competition, Task 1. Task 1 targets the automatic classification of tweets that mention medication. We adapted the standard BERT pretrain-then-fine-tune approach to include an intermediate training stage with a biLSTM architecture neural network acting as a further fine-tuning stage. We were inspired by the effectiveness of within-task further pre-training and sentence encoders. We show that this approach works well for a highly imbalanced dataset. In this case, the positive class is only 0.2% of the entire dataset. Our model performed better in both F1 and precision scores compared to the mean score for all participants in the competition and had a competitive recall score.

1 Introduction

Social media is ubiquitous, a continuous part of our daily lives; it offers new ways of communication and contains important data for various disciplines (Pershad et al., 2018). This is especially crucial for health related sectors. Twitter posts are now recognised as an important source of patient-generated data, providing unique insights into population health (Nguyen et al., 2017). However, tweets contain a lot of noise and are imbalanced in their nature. A fundamental step towards incorporating Twitter data in pharmacoepidemiologic research is to automatically recognise medication mentions in tweets. Hence Task 1 in SMM4H 2020 (Ari Z. Klein and Gonzalez-Hernandez, 2020) is proposed based on a similar task from 2018 (Weissenbacher et al., 2018) to address these challenges. The task is aimed at identifying if tweets contain information about medications using a highly skewed, imbalanced training dataset in which only 0.2% contains a positive label.

2 Methodology

2.1 Data

The *training data* provided by SMM4H 2020 consists of 69,272 training data, out of which only 181 are positive tweets, and 69,091 are negative tweets. I.e., it contains binary labelled unique tweets that mention a medication or dietary supplement (annotated as "1") and tweets that do not (annotated as "0"). Compared with the same task in 2018, where artificially balanced tweet data was provided, the current training data consist of highly imbalanced tweets, in which only 0.2% mentioned a medication (annotated as "1"), corresponding to a more realistic real-world scenario. The *test data* provided contains 29,687 examples. Additionally, the previously mentioned artificially balanced tweet dataset used for the SMM4H 2018 Task 1 (Weissenbacher et al., 2018), was released to participants, containing 2,367 unique tweets with 1,212 positive and 1,155 negative labels.

2.2 Data Cleaning and Pre-processing

As the distinct way the reference to medication is made is not important in this binary classification tasks, we applied various standard pre-processing and cleaning methods on the tweets, as follows. Be-

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 165–167
Barcelona, Spain (Online), December 12, 2020.

fore training, we processed the datasets using the Ekphrasis library (Baziotis et al., 2017). The library provides text preprocessing operations optimised for social media texts. Each tweet was normalised, by replacing all Twitter usernames and URLs with common tokens, such as <*user*>and <*url*>. Hashtags, emoticons, elongated and repeated words were further annotated with special tokens using the 'Social Tokenizer' provided in the above library. As an extra preprocessing step, we used spell correction and word segmentation tools from the same library to correct misspelled words and split long texts into more meaningful sub word forms.

2.3 Preliminary Fine-tuning of the Model on a Balanced Task

Before training with the SMM4H 2020 Task 1 training data, we use the artificially balanced data from 2018 to fine-tune our model; here, we apply BERT (Devlin et al., 2018), as the most advanced training network available currently, with a learning rate of 2e-5. As BERT is not a classifier, following the fine-tuning process (Devlin et al., 2018; Sun et al., 2019), we use the last layer of BERT and build a linear layer on top of it, as a classification task. We perform this preliminary fine-tuning as we expect that the balanced dataset allows the BERT model to better understand and capture the generality of the data distribution for tweets that mention medications.

2.4 Target Task Training with BERT Feature Information

After the preliminary training of BERT, the model can be considered as containing the contextual information of medication tweets data. Inspired by the effectiveness of the within-task further pre-training and sentence encoder (Phang et al., 2018; Sun et al., 2019), we further experimented with the idea of a supplementary training phase on a data-rich supervised task. This process involves training BERT using the output as features in our target task. We experimented with different outputs from the last hidden states. As recommended in the original BERT paper (Devlin et al., 2018), we experimented with the feature-based approach using the activations from the last hidden layer, the sum of all 12 hidden layers, the sum of the last four hidden layers and the concatenation of the last four hidden layers as input into our classification model.

The model parameters learned from the intermediate task served as contextualised sentence embeddings (Phang et al., 2018) to initialise our target task model. We feed these embedding to a bidirectional LSTM (biLSTM) model (Graves and Schmidhuber, 2005) to classify the tweets. We experimented with the output of each encoder layer, to see which vector provides the best contextualised embedding. These experiments showed that the sum of the final four hidden layers of the pre-trained model provides the best embedding. We then trained the biLSTM model for 5 epochs to reach optimal performance using a binary cross-entropy loss.

For the intermediate training, we utilise the features generated by BERT as the contextual sentence embedding, similar to word embedding. We compared the difference in performance by freezing and unfreezing the feature vectors from BERT. This comparison suggest that unfreezing achieves better performance. This can be considered as a transfer learning phase which results in a boost of performance.

3 Results

As mentioned previously, the information from BERT can be considered as a sentence embedding to the supplementary BiLSTM model; we have the option to either freeze or unfreeze the BERT model parameters in the target task training. Freezing the parameters means that we use a fixed prior knowledge from the balanced dataset as a sentence embedding to the BiLSTM model, while unfreezing the parameters allows the parameters in the sentence embedding to update with the new data, similarly to how the NLP field uses pre-trained word embeddings and then fine-tunes them with new data. The results from the experiments, when we freeze and unfreeze the BERT part, are presented in Table 1 and Table 2, respectively. All the results in both tables are calculated based on the unbalanced data. As said, the best performing model on the imbalanced dataset results from concatenating the final four layers. Our system achieved an F1 score on the positive (minority) class of 0.7164, with 0.8421 precision and 0.6234 recall on the test dataset. The mean score for this task was an F1 score on the positive class of 0.6646, with

0.7007 precision and 0.7039 recall.

4 Conclusion

The results of our experiments suggest that extracting features from language models and using them for intermediate modelling can provide a good platform to study the automatic classification of medical tweets, even if the final target dataset is highly imbalanced. Moreover, interestingly, various layers of features from language models can be used as a sequence in the representation, and this further confirms that different layers of representation in complex language models, such as BERT, contain various levels of information semantics.

Layers	F1	Precision	Recall	TP/FP/FN
concatenate last 4 hidden layers	0.78	0.79	0.77	27/7/8
last hidden layer*	**0.81**	**0.84**	**0.77**	**27/5/8**
sum of last 4 hidden layers	0.75	0.78	0.71	25/7/10
second to last hidden layers	0.81	0.84	0.77	27/5/8

Table 1: Experiments results for freezing parameters of BERT

Layers	F1	Precision	Recall	TP/FP/FN
concatenate last 4 hidden layers*	**0.88**	**0.97**	0.80	28/1/7
last hidden layer	0.83	0.81	**0.86**	**30/7/5**
sum of last 4 hidden layers	0.85	0.83	**0.86**	30/6/5
second to last hidden layer	0.84	0.85	0.83	29/5/6

Table 2: Experiments results for not freezing parameters of BERT

References

Ivan Flores Arjun Magge Zulfat Miftahutdinov Anne-Lyse Minard Karen O'Connor Abeed Sarker Elena Tutubalina Davy Weissenbacher Ari Z. Klein, Ilseyar Alimova and Graciela Gonzalez-Hernandez. 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 (#SMM4H) Workshop & Shared Task*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

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*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Quynh C Nguyen, Kimberly D Brunisholz, Weijun Yu, Matt McCullough, Heidi A Hanson, Michelle L Litchman, Feifei Li, Yuan Wan, James A VanDerslice, Ming Wen, et al. 2017. Twitter-derived neighborhood characteristics associated with obesity and diabetes. *Scientific reports*, 7(1):1–10.

Yash Pershad, Patrick Hangge, Hassan Albadawi, and Rahmi Oklu. 2018. Social medicine: Twitter in healthcare. *Journal of Clinical Medicine*, 7(6):121, May.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Davy Weissenbacher, Abeed Sarker, Michael Paul, and Graciela Gonzalez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 13–16.