John_Snow_Labs@SMM4H'22: Social Media Mining for Health (#SMM4H) with Spark NLP

Veysel Kocaman[1], Cabir Celik[1], Damla Gurbaz[1], Gursev Pirge[1], Bunyamin Polat[1], Halil Saglamlar[1], Meryem Vildan Sarikaya[1], Gokhan Turer[1], David Talby[1]

[1]John Snow Labs Inc., 16192 Coastal Highway, Lewes, DE , USA 19958,
veysel@johnsnowlabs.com

Abstract

Social media has become a major source of information for healthcare professionals but due to the growing volume of data in unstructured format, analyzing these resources accurately has become a challenge. In this study, we trained health related NER and text classification models on different datasets published within the Social Media Mining for Health Applications (#SMM4H 2022) workshop[1]. We utilized transformer based Bert for Token Classification and Bert for Sequence Classification algorithms as well as vanilla NER and text classification algorithms from Spark NLP library during this study without changing the underlying DL architecture. The trained models are available within a production-grade code base as part of the Spark NLP library; can scale up for training and inference in any Spark cluster; has GPU support and libraries for popular programming languages such as Python, R, Scala and Java.

1 Introduction

A primary building block in text mining systems is Named Entity Recognition (NER) - which is regarded as a critical precursor for question answering, topic modeling, information retrieval, etc. (Li et al., 2022). In the medical domain, NER recognizes the first meaningful chunks out of a clinical note, which are then fed down the processing pipeline as an input to subsequent downstream tasks such as clinical assertion status detection (Uzuner et al., 2011), clinical entity resolution (Tzitzivacos, 2007) and de-identification of sensitive data (Uzuner et al., 2007).

Text classification is also one of the important tasks of machine learning and has been extensively used in several areas of NLP. It can be defined as assigning a sentence or document an appropriate category.

In this paper, we share our results on the #SMM4H Shared Task, which involves NLP challenges of using social media data for health research, including informal, colloquial expressions and misspellings of clinical concepts, noise, data sparsity, ambiguity, and multilingual posts. The contest involves classification and NER tasks. As John Snow Labs team, we contributed to many tasks and trained several NER and Classification models on English and Spanish social media data.

2 Background

2.1 Spark NLP Library

All the experiments are run within Spark NLP library, a popular open-source NLP library that has been downloaded more than 30 million times so far[2].

Spark NLP library has two versions: Open source and enterprise. The open-source version has all the features and components that could be expected from any NLP library, using the latest Deep Learning (DL) frameworks and research trends. Enterprise library is licensed (free for academic purposes) and designed towards solving

[1] https://healthlanguageprocessing.org/smm4h-2022/

[2] https://pepy.tech/project/spark-nlp

Proceedings of the 29th International Conference on Computational Linguistic, pages 44–47
October 12–17, 2022.

real world problems in the healthcare domain and extends the open-source version. The licensed version has the following modules to help researchers and data practitioners in various means: Named entity recognition (NER), assertion status (negativity scope) detection, relation extraction, entity resolution (SNOMED, RxNorm, ICD10 etc.), clinical spell checking, contextual parser, deidentification and obfuscation (Kocaman and Talby, 2021).

2.2 Named Entity Recognition in Spark NLP

There are two DL architectures for NER tasks in Spark NLP: *BiLSTM-CNN-Char* and *BertForTokenClassification. BiLSTM-CNN-Char* architecture (Kocaman and Talby, 2020) is a modified version of the architecture proposed by Chiu and Nichols (2016). It may be defined as a neural network architecture that automatically detects word and character-level features using a hybrid bidirectional LSTM and CNN architecture, eliminating the need for most feature engineering steps.

NER systems are usually a part of an end-to-end NLP pipeline through which the text is fed and then several text preprocessing steps are applied. Since the DL algorithm we implement is sentence-wise, and the features (embeddings and casing) are token-wise, sentence splitting and tokenization are the most important steps leading to better accuracy. Using a DL based sentence detector module (Scweter and Ahmed, 2019) and a highly customizable rule-based tokenizer in Spark NLP, we ensured that the generated features are more informative.

The second architecture, *BertForTokenClassification (BFTC)*, can load Bert (Peters et al., 2018; Radford et al., 2018) based models with a token classification head on top (a linear layer on top of the hidden-states output) e.g., for NER tasks. Spark NLP lets users utilize this transformer architecture into Spark NLP with just a few lines of code.

2.3 Classification in Spark NLP

To be able to process large volume of data, the text classification model needs to be scalable, and accurate, as it is used to filter out documents, reviews, and tweets that do not contain any indication of adverse events. To achieve this, Spark NLP offers two DL architectures: *ClassifierDL* and *Bert for Sequence Classifier* (BFSC). ClassifierDL uses a Fully Connected Neural Network (FCNN) model that does not require hand-crafted features and relies on a single embedding vector for classification. Given the conversational nature of social media text, we can utilize the entire document to get efficient embeddings (with little text clipping in case of BioBERT embeddings) that we directly feed to the classifier model. Since there is only a single feature vector as input to the model, we test multiple embedding techniques to analyze performance (Haq et al., 2022).

Similar to the token classification problem for NER tasks, *BertForSequenceClassification (BFSC)* is used for text classification tasks based on Bert architecture.

3 Implementation Details

Models for classification tasks were trained by using Bert for Sequence Classifier (BFSC) and Bert for Token Classifier (BTFC) was used for training NER tasks. Embeddings and hyper parameters used for training the documents are shown in Table 1. Once the pre-trained models were loaded within a pipeline, those models were used for prediction. *MedicalBertForTokenClassifier* and *MedicalBertForSequenceClassification* annotators were used to test the performances of the models on the validation datasets.

4 Experimental Results

Using the official training sets from the contest, we trained models and obtained metrics on the test sets used in the challenges. The results are shown in Table 1, where all the tested configurations obtained decent accuracies over the validation dataset. Tasks 1,2,3a, 5 and 10 are not worked on as a team; but some team members made separate individual efforts on them.

Unless otherwise specified, Biobert-base-cased-v1.2 embeddings were used for model training of English texts. Often times, cleaning the social media material (removing emojis, tags, repeating punctuations etc.) did not improve the results.

Task 1a is about classifying tweets reporting Adverse Drug Events (ADEs). BFSC gave the best results. The training dataset was unbalanced; therefore, it has been enriched with different ADE datasets (Haq et al., 2022) to increase the number of ADE entities.

Task 1b involves detecting ADE intervals in tweets. The model was trained using BFTC. As mentioned above, the training dataset was unbalanced and just like Task 1a, the dataset was enriched by using datasets (Haq et al., 2022) with the aim of increasing the number of ADE entities.

Table 1. Detailed information about the competition tasks, embeddings, parameters and model performance.

Task	Additional Data	Model Performance					
		P	R	F1	P	R	F1
1a	Yes	Validation			Test		
		0.97	0.96	0.96	*		
1b	Yes	Validation			Test		
		0.89	0.88	0.88	*		
2a	No	Validation			Test		
		0.72	0.72	0.72	*		
2b	No	Validation			Test		
		0.74	0.72	0.73	*		
3a	No	Validation			Test		
		0.79	0.72	0.75	*		
3b	No	Validation			Test		
		0.86	0.85	0.85	0.84	0.82	0.83
4	No	Validation			Test		
		0.79	0.85	0.82	0.82	0.80	0.81
5	No	Validation			Test		
		0.77	0.77	0.77	*		
6	Yes	Validation			Test		
		0.79	0.78	0.78			
7	Yes	Validation			Test		
		0.80	0.66	0.73	0.88	0.66	0.71
8	No	Validation			Test		
		0.72	0.86	0.78	0.68	0.88	0.76
9	No	Validation			Test		
		0.85	0.90	0.87	0.86	0.85	0.86
10	No	Validation			Test		
		0.69	0.87	0.77	*		

***** Tasks 1,2,3a, 5 and 10 are not worked on as a team; but some of the team members did some individual efforts on the side.

In Task 2a, classification of stance in tweets about health mandates is expected. BFSC with biobert-base-cased-v1.2 embeddings produced the best results, whereas bert_base_cased embeddings produced slightly higher metrics.

Task 2b involves binary classification of premise in tweets about health mandates. BFSC produced the better F1 score values.

In Task 3a, detection of tweets where the users self-declare changing their medication treatments is expected. The model trained using BFSC produced higher F1 score values. Training data was unbalanced and using different embeddings and cleaning the tweets did not improve the results.

Task 3b involves binary classification of drug reviews from WebMD.com. BFSC produced around 0.84 F1 score values.

Task 4 involves identification of the self-report exact age in tweet posts. BFSC produced the best results (F1-score of 0.82).

For Task 5, all the tested configurations obtained decent accuracies over the validation dataset. The difference of scores between the two classes is probably due to the imbalance in the dataset.

Task 6 involves identification of self-reported COVID-19 vaccination status in English tweets. The two classes in the provided training dataset are Vaccine_chatter and Self_reports (tweets of users clearly stating that they have been vaccinated). The classes are significantly imbalanced with a ratio of 1 to 8. BFSC produced an F1 score value (for the positive class) of 0.72.

Task 7 involves the identification of self-reported Intimate partner violence (IPV) in English tweets. The dataset is significantly unbalanced, and the negative tweets include non-IPV domestic violence and non-self-reported IPV, which can hardly be distinguished. Where the model could not select between domestic violence and IPV, better results were obtained by introducing family-related words (father, mother, daughter, sister etc.) to the model and enabling it to learn about domestic violence. BFSC produced F1 values around 0.72 (for the positive class) for the validation dataset.

Task 8 is about detecting tweets that are self-disclosures of chronic stress. For the classification task, a pre-trained Roberta For Sequence Classification model from Hugging Face with Roberta embeddings is fine-tuned on the given dataset. Without cleaning the tweets, the model resulted in better scores. The validation F1 score was 0.78 and the test F1 score is 0.76.

Task 9 involves identification of the exact age in social media forum (Reddit) posts. BFSC produced the best results (F1-score of 0.87).

Task 10 requires detection of disease mentions in Spanish tweets, and BFTC was used to produce the models. Huggingface's 5-language embeddings (bert-base-5lang-cased) were used to get the highest F1 score values.

5 Conclusion

In this study, we show through experiments on #SMM4H contest datasets in two languages that models trained by BFSC and BFTC can easily be adapted to the Spark NLP environment and achieve decent scores on social/health-related datasets with zero code changes. We trained BFSC models for classification and BFTC models for NER purposes that can be used as a pretrained model out of the box within Spark NLP for Healthcare library.

Considering the F1 scores, all the models produced decent performances during training. The problem of getting low metrics due to unbalanced datasets was solved by enriching the dataset with data from similar datasets and consequently, there was a substantial increase in the F1 scores.

References

Jason P. C. Chiu and Eric Nichols. 2016. *Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics*. 4 (2016) 357–370. https://doi.org/10.48550/arXiv.1511.08308

Hasham Ul Haq, Veysel Kocamanand David Talby. 2022. Mining adverse drug reactions from unstructured mediums at scale. W3PHIAI workshop at AAAI-22. https://doi.org/10.48550/arXiv.2201.01405.

Veysel Kocaman and David Talby. 2020. *Improving Clinical Document Understanding on COVID-19 Research with Spark NLP.* arXiv:2012.04005v1

Veysel Kocaman and David Talby. 2021. *Spark NLP: Natural Language Understanding at Scale. Software Impacts*, arXiv:2101.10848v1 [cs.CL].

Jing Li, Aixin Sun, Jianglei Han and Chenliang Li. 2022. *A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering,* 34(1). https://doi.org/10.48550/arXiv.1812.09449.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. *Deep contextualized word representations*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, page 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. *Improving language understanding with unsupervised learning*. Technical report, OpenAI.

Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-purpose neural networks for sentence boundary detection. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*.

Dimitri Tzitzivacos. 2007. *International Classification of Diseases 10th edition (ICD-10): main article*. *CME: Your SA Journal of CPD.* 25(1): 8–10.

Ozlem Uzuner, Yuan Luo and Peter Szolovits. 2007. *Evaluating the State-of-the-Art in Automatic De-identification*. *Journal of the American Medical Informatics Association*. 14(5): 550–563

Ozlem Uzuner, Brett South, Shuying Shen and Scott L. DuVall. 2011. *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association* 18(5): 552–556