Edinburgh_UCL_Health@SMM4H'22: From Glove to Flair for handling imbalanced healthcare corpora related to Adverse Drug Events, Change in medication and self-reporting vaccination

Imane Guellil
University of Edinburgh

Jinge Wu
University College London

Honghan Wu
University College London

Tony Sun
University of Waterloo

Beatrice Alex
University of Edinburgh

Abstract

This paper reports on the performance of Edinburgh_UCL_Health's models in the Social Media Mining for Health (SMM4H) 2022 shared tasks. Our team participated in the tasks related to the Identification of Adverse Drug Events (ADEs), the classification of change in medication (change-med) and the classification of self-report of vaccination (self-vaccine). Our best performing models are based on DeepADEMiner (with respective F1= 0.64, 0.62 and 0.39 for ADE identification), on a GloVe model trained on Twitter (with F1=0.11 for the change-med) and finally on a stack embedding including a layer of Glove embedding and two layers of Flair embedding (with F1= 0.77 for self-report).

1 Introduction

Identification of healthcare-related topics from social media is considered meaningful work in society. Therefore, it is attracting attention, particularly from pharmaceutical companies who want to know what people (i.e. patients) think and report about their products. Technically, it requires the ability to apply Natural Language Processing (NLP) techniques for the automatic collection, extraction, representation, analysis and validation of social media data such as Twitter and Reddit posts. This paper summarises our work in the Social Media Mining for Health Applications (#SMM4H) workshop (Davy Weissenbacher, 2022), including identification of ADEs (Task 1), classification of change in medication regimen in tweets (Task 2) and classification of tweets indicating self-reported COVID-19 vaccination. The first task splits into three further sub-tasks of Classification of English tweets reporting ADEs (Task 1a), Extraction of ADE spans in tweets (Task 1b) and Normalization of these colloquial mentions to their standard concept identifiers (IDs) in the MedDRA vocabulary (Task 1c) (Mozzicato, 2009). All of these are explained in detail in Section 2.

2 Task 1: Classification, extraction and normalization of adverse effect mentions

2.1 Data and Pre-processing

The dataset (provided by SMM4H's organizers) includes three parts:

- A training set consisting of 17,385 tweets with 1,711 ADE examples (1,240 examples without duplicates) and 16,145 examples without ADEs (NoADEs).
- A validation dataset containing 915 tweets with 87 ADE examples (65 without repetition) and 850 NoADE examples.
- A test dataset including 10,984 tweets.

The corpus is highly imbalanced as only 14% of the tweets contain ADEs.

We use the *tweet-preprocessor* python API [1] to replace the ambiguous mentions (e.g., typos in sentences) by correcting words and removing URLs and emojis. However, we only use this type of pre-processing for the classification using the first model (Glove_FlairFor_FlairBack) that we detail in Section 2.2.1.

2.2 Sub-task 1a: ADE tweet classification

This sub-task aims to detect tweets that contain an adverse effect (AE), also known as adverse drug effect (ADE), and label them with the tag ADE.

2.2.1 Method

For this binary classification task (containing ADE or not), we compare the performance of two main embeddings, transformers (mainly represented by BERT, RoBERTa) and contextual (mainly represented by FLAIR).

[1] https://pypi.org/project/tweet-preprocessor/

Proceedings of the 29th International Conference on Computational Linguistic, pages 148–152
October 12–17, 2022.

For this purpose, we use two models: 1) Glove_FlairFor_FlairBack, a stack embedding including three types of embeddings (GloVe (Pennington et al., 2014), Flair-Forward and Flair-Backward (Akbik et al., 2018)) and 2) DeepADEMiner_default, a BERT-based model (classifier-bertweet-large) trained on the training data of the shared task , (Magge et al., 2021).[2]

For training the first model (Glove_FlairFor_FlairBack), we use the FLAIR NLP framework (*which enables model training for sequence-labelling-based text classification*). (Akbik et al., 2019). A Long short-term memory network (LSTM) (Hochreiter and Schmidhuber, 1997) was also used over the word embedding in a document (i.e., each tweet) to output the document representation. Document embeddings are different from word embeddings in that they compute one embedding for an entire text, whereas word embeddings produce embeddings for individual words. The Glove_FlairFor_FlairBack model was trained for 30 epochs with a learning rate of 0.05 and the hidden_size variable set to 256. For handling the imbalance in the data, the weight of the ADE label was set to 10, whereas the one for NoADE was kept at 1.

For the second model (classifier-bertweet-large) we used the framework DeepADEMiner and in particular the default model. For training, we also used the embeddings of the large Roberta model *roberta-large*[3]. The DeepADEMiner_default model was trained for 10 epochs with a learning rate of 0.000001. To deal with the imbalanced corpus, the weight loss was fixed to 10 for the ADE class. The corpus was also randomly under-sampled retaining only 20% of the tweets with the label of NoADE. In terms of the evaluation, we directly applied the original model (DeepADEMiner_default) for evaluation without further fine-tuning due to time constraints[4].

2.2.2 Results

During the validation phase, we only trained the Glove_FlairFor_FlairBack model. In the second model (DeepADEMiner_default), we used the default trained model directly on the test dataset. The results regarding both stages (development and testing) are presented in Table 1. We also included the mean results related to the mean score calculated from the results returned by all participants.

Figure 1: The classification results on the dev dataset

Corpus	Model	P	R	F1
Dev	Glove_FlairFor_FlairBack	0.4804	0.7538	0.5868
Test	Glove_FlairFor_FlairBack	0.452	0.505	0.477
	DeepADEMiner_default	0.554	**0.765**	**0.642**
	Mean_results	**0.646**	0.497	0.562

2.2.3 Analysis

By increasing the weight loss to 10 (based on the previous works proposed by (Magge et al., 2021)), we observed that both models classified many false positives. The majority of the tweets that were wrongly classified as ADE include only the name of drugs without any ADEs. Some of them include the reference to *pain or a burn on the skin* which were the initial symptoms and not ADEs related to drugs. In future works, we plan to investigate the relation between weight loss and the results.

2.3 Sub-task 1b: ADE span extraction

2.3.1 Method

We first explored two methods for the span extraction task: 1) The default model used in DeepADEMiner (latest version using *roberta-large*[5]) and 2) A combination of results returned by the default model (*roberta-large*) and a flair-forward model trained using the DeepADEMiner framework.

For the first model, DeepADEMiner_default, we used the default parameters training the model for 10 epochs with a learning rate of 0.000001 and the hidden_size setting of 256. Only 5% of the corpus was used for validating this model.

For the second method, DeepADEMiner_default_FlairFor, we use the union of the two model outputs and we remove any duplicates. The flair-forward model was trained using 100 epochs with all other parameters kept as their defaults. The unique difference is that we used 10% of the corpus for validating this combined model. As a post-processing step, punctuation was automatically removed in case it was detected as an ADE for both approaches.

2.3.2 Results

Table 2 presents the results obtained using DeepADEMiner_default and DeepADEMiner_default_FlairFor on the test dataset. It

[2] https://bitbucket.org/pennhlp/deepademiner/src/master/

[3] https://huggingface.co/roberta-large

[4] https://hlp.ibi.upenn.edu/public_downloads/DeepADEMiner/model/latest/classifier/

[5] https://huggingface.co/roberta-large

also presents the mean scores representing the overlapping and the strict score related to the n precision, recall and f1 score.

Figure 2: The extraction results on the test dataset where _over is for the overlapping results and $_{s}trisstrictresults$

Model	P_over	R_over	F1_over	P_str	R_str	F1_str
DeepADEMiner_default	**0.569**	0.691	**0.624**	**0.427**	**0.526**	**0.471**
DeepADEMiner_default_FlairFor	0.531	**0.709**	0.607	0.352	0.484	0.408
Mean_results	0.539	0.517	0.527	0.344	0.339	0.341

2.3.3 Analysis

Based on the results, we observe that overall DeepADEMiner_default achieves the best performance. DeepADEMiner_default_FlairFor also performs well, especially on the R_over. SemEHR seems to be poor in this task. It has a huge number of false positives. The reason for this is that it is less powerful when facing ambiguous mentions (e.g., "triggers my rapid cycling") and informal language (e.g., "can't sleeeep"). However, the deep learning models could get rid of this due to contextual learning from words. Also, SemEHR isn't designed to just detect ADEs. It detects all types of medical terms. It is mainly for this reason that it performed poorly on our in-domain data.

2.4 Sub-task 1c: ADE resolution

2.4.1 Method

The main goal of the ADE resolution task is to assign each ADE mention found in the text to its corresponding MedDRA code (Link). The file generated during the extraction phase in Sub-task 1b using the best-performing model (i.e., DeepADEMiner_default) was used as input into the resolution sub-task. Different approaches were considered for resolution but the two best-performing models selected were:

1) the default DeepADEMiner model (based on the model *bert-base-uncased*[6], and

2) DeepADEMiner_default+mcn-en-smm4h a pre-trained hugging face model called mcn-en-smm4h was applied[7]. This model was pre-trained using BioBERT and smm4h 2017 data (Sarker et al., 2018). This model was then fine-tuned using a hidden_size of 768, a max_position_embeddings of 512, an attention_probs_dropout_prob of 0.1, and hidden_dropout_prob of 0.1. For each mention, it outputs the linked MedDRA term with the preferred term (PT)/lowest level term (LLT) identifier.

2.4.2 Results

Table 3 lists the results obtained by using the two models just described. One is the default model used in the DeepADEMiner system (DeepADEMiner_default). The other model DeepADEMiner_test+mcn-en-smm4h refers to the normalization results fine-tuned with the *mcn-en-smm4h* model based on the Sub-task 1b results. It also presents the mean scores representing the overlapping and the strict score related to the precision, recall and f1 score.

Figure 3: The normalization results on the test dataset

Model	P_over	R_over	F1_over	P_str	R_str	F1_str
DeepADEMiner_default	0.04	0.04	0.04	0.03	0.03	0.03
DeepADEMiner_default+mcn-en-smm4h	**0.35**	**0.43**	**0.39**	**0.29**	**0.35**	**0.32**
Mean_results	0.120	0.112	0.116	0.085	0.082	0.083

2.4.3 Analysis

For the test dataset, DeepADEMiner_default+mcn-en-smm4h achieves the best performance with an F1_over score of 0.387. This may appear relatively low but presumably, this true positive linked entity can only be achieved if it was extracted properly in the first place. So the resolution performance is affected by the extraction one.

3 Task 3a: Classification of change in medication regimen in tweets

3.1 Data and pre-processing

The organizers provided the dataset of tweets where users self-declare changing their medication treatments, regardless of being advised by a health care professional to do so. This dataset includes 5,898 Tweets for training, 1,572 Tweets for validation and 15,360 for testing, the majority of them being decoy Tweets as only 2,360 Tweets were considered by the organizers for final system evaluation. This dataset is highly imbalanced as only 519 and 139 tweets in the training and the validation datasets, respectively, self-report a change in medication, representing only 9% of the whole dataset. No pre-processing steps were applied to this dataset.

3.2 Method

For this sub-task, we use the Glove Twitter embedding model (pre-trained glove vectors based on 2B tweets, 27B tokens, 1.2M vocab, uncased)[8]. We use the default Glove_Twitter model provided by Flair NLP.The tweets were kept in their original

[6] https://huggingface.co/bert-base-uncased

[7] https://huggingface.co/olastor/mcn-en-smm4h

[8] https://nlp.stanford.edu/projects/glove

form. As we did for classifying ADEs, we also used the flair NLP framework for training (Akbik et al., 2019). Same for the ADE classification, LSTM was also used over the word embedding in a document to output the document representation. The model was trained for 30 epochs with a learning rate of 0.05 and a hidden_size setting of 256. For handling the imbalance, the weight of tweets containing a change in medication (with the label 1) was set to 10 whereas the weights of other tweets were kept at 1.

3.3 Results

Table 4 shows the results of the Glove_Twitter model obtained on the development and test datasets. It also presents the mean scores of all participants.

Figure 4: The change of medication results on the validation dataset

Corpus	Model	P	R	F1
Dev	Glove_Twitter	0.34	0.30	0.32
Test	Glove_Twitter	**0.54**	0.11	0.19
	Mean_results	0.46	**0.54**	**0.46**

3.4 Analysis

The majority of errors are related to the misclassification of tweets including some medications that were prescribed for the first time. To improve the results, the training corpus should be enriched with more samples related to the first-time medication. Also, the chosen model is a fast model to run, however, it is not the best model for taking into account the context. Hence, we plan to use contextual embedding including Flair (Akbik et al., 2019) or Elmo (Peters et al., 2018), in the future, for improving performance.

4 Task 6: Classification of tweets indicating self-reported COVID-19 vaccination

4.1 Data and pre-processing

The shared task organisers provided the dataset of tweets where users self-declare their COVID-19vaccination status. This dataset includes 13,693 tweets for training, 2,784 tweets for development and 5,923 tweets for testing.

4.2 Method

For this sub-task, we use two main models: 1) a Glove Twitter embedding model (the same that we use for Task 3), and 2) a stack embedding including three types of embeddings including Glove, Flair-Forward and Flair-Backward (the same which we use for Sub-task 1a).

With the first, the Glove_Twitter, model the tweets were kept in their original form. As we did for classifying ADEs. This model was trained for 26 epochs. In contrast, we tried three different epochs for training the second, stack embedding model, i.e., 6, 15 and 30 epochs. For training all the models, we also used the flair NLP framework (Akbik et al., 2019). As for the previous tasks, LSTM was also used over the word embedding in a document to output the document representation. Both models were trained with a learning rate of 0.05 and a hidden_size setting of 256. No pre-processing, sampling or increase in weights was applied to the first model. However, a random under-sampling method, where we kept the same number of tweets for both classes, was applied for the second model when the model was trained for 15 epochs. The tweets were also prepossessed using the Twitter-preprocess API when the model was run for 6 and 30 epochs. Also, the weight of tweets self-reporting a vaccination (with label 1) was set to 3 (where the models were trained for 6 and 30 epochs) and 10 (where the model was trained for 15 epochs) whereas the weights of all other tweets were kept 1.

4.3 Results

Table 5 illustrates the results obtained on the test dataset. It also presents the median scores of all participants.

Figure 5: The vaccine self-reporting results on the test dataset

Model	P	R	F1
Glove_Twitter	0.87	0.57	0.69
Glove_FlairFor_FlairBack_epoch6	0.67	0.80	0.73
Glove_FlairFor_FlairBack_epoch15	0.47	**0.93**	0.63
Glove_FlairFor_FlairBack_epoch30	0.69	0.86	**0.77**
Median_results	**0.9**	0.77	0.68

4.4 Analysis

As the majority of people reporting a vaccination are using other terms such as "COVID-19", when these terms are used in a vaccine chatter context, they are misclassified. Also, the tweets that are not directly referencing vaccination or the sarcastic tweets are misclassified.

5 Conclusion

In this work, we explore the use of different models associated with different techniques for dealing with imbalanced healthcare-related social media corpora. We observed that the best technique relied on the use of deep learning models such as *Roberta or Flair*, under-sampling and the adjustment of the weight loss. In the future, we plan to explore more techniques for handling imbalanced datasets, augmenting the training corpus and relying on more embedding models.

6 Acknowledgement

This study/project is funded by the National Institute for Health Research (NIHR) Artificial Intelligence and Multimorbidity: Clustering in Individuals, Space and Clinical Context (AIM-CISC) grant NIHR202639. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Luis Gascó Darryl Estrada-Zavala Martin Krallinger Yuting Guo Yao Ge Abeed Sarker Ana Lucia Schmidt Raul Rodriguez-Esteban Mathias Leddin Arjun Magge Juan M. Banda Vera Davydova Elena Tutubalina Graciela Gonzalez-Hernandez Davy Weissenbacher, Ari Z. Klein. 2022. Overview of the seventh social media mining for health applications smm4h shared tasks at coling 2022. *In Proceedings of the Seventh Social Media Mining for Health (SMM4H) Workshop and Shared Task.*

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Deepademiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Patricia Mozzicato. 2009. Meddra. *Pharmaceutical Medicine*, 23(2):65–75.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.