Adverse Drug Reaction Classification in Social Media: A Multi-label Approach

Julie Durand*†, Athena Stassopoulou*, Ioannis Katakis*
* Department of Computer Science, School of Sciences and Engineering, University of Nicosia, 2417 Nicosia, Cyprus
† Pharmacovigilance Office, European Medicines Agency, Amsterdam, The Netherlands
email: durand.j@live.unic.ac.cy, stassopoulou.a@unic.ac.cy, katakis.i@unic.ac.cy

***Abstract*—Many patients readily share experiences about their medical conditions and treatments on online social media, which makes these platforms a potentially valuable source of information on adverse drug reactions (ADRs). In this work, the detection of mentions of ADRs in Reddit posts is approached as a multi-label classification problem. A dataset of 537 annotated posts was created by supplementing a publicly available dataset with freshly collected and annotated posts. The labels were mapped to the Medical Dictionary for Regulatory Activities (MedDRA) and their distribution within each MedDRA level guided the creation of 12 data subsets. On each data subset, we applied 4 different multi-label learning methods – Binary Relevance (BR), Classifier Chains (CC), Label Powerset (LP) and random k-labelsets (RAkEL), each associated with 4 different base classifiers: Decision Trees (DT), Naïve Bayes (NB), Random Forest (RF) and Support Vector Machine (SVM). The best F-scores were with DT on the data subset based on the 20 most frequent labels at MedDRA Preferred Term (PT) level. The best hamming loss was with the data subset based on all labels at PT level. The type of multi-label learning method did not appear to influence performance significantly. Our results show a promising direction in the use of multi-label classification of ADRs from social media posts for pharmacovigilance purposes.**

***Index Terms*—multi-label classification, adverse drug reaction, social media**

I. Introduction

A. Background

An adverse event (AE) can be defined as any undesirable medical occurrence in a patient administered a medicine. Although strictly speaking, an adverse drug reaction (ADR) refers to an AE for which there is at least a reasonable possibility of a causal association with a drug, we use the term ADR throughout this paper regardless of causality to be consistent with the published literature on social data mining using machine learning. ADRs are associated with an important morbidity and mortality and put a significant burden on healthcare systems worldwide. Their early detection and characterisation are therefore crucial and are objectives of pharmacovigilance [1]. Spontaneous reporting systems, a pillar of traditional pharmacovigilance, hold structured information on suspected ADRs reported by healthcare professionals and patients, but they are subject to several limitations, including under-reporting or delayed reporting.

Many patients readily share their experience about their medical conditions and treatments on online social media, which has stimulated research on methods to harness data from social media for pharmacovigilance purposes [2, 3, 4, 5]. Social media mining however comes with multiple challenges, including high data volume, incomplete information, false information, casual expression of ADRs, therapeutic indications mistaken for ADRs and high data imbalance [4, 6, 7, 8].

B. Prior Work

At over 500 million posts per day, Twitter is one of the most popular social media platforms among both users and researchers [6, 8, 9]. Other online data sources used by researchers include DailyStrength [2, 3, 7] or Facebook [5]. Due to the imbalanced nature of the data, large numbers of posts have to be annotated to establish the ground truth. Research teams usually employ two or three independent domain experts for the task and resolve any disagreement by consensus reconciliation or majority opinion [2, 3, 4, 5, 7].

Beyond n-grams, a common way to turn pre-processed posts into model features is to use word embeddings, as fixed constants or learnable parameters [3, 4, 6, 10, 11, 12]. One of the earliest attempts to extract ADRs from social media relied on lexicon terms that were retrieved from posts using a sliding window [2]. Support vector machines (SVMs) have been shown to perform well in text classification tasks and have been used to classify ADRs [5, 7, 12] or determine whether an ADR was related to a drug [8]. Other machine learning approaches explored include conditional random field (CRF) [3], Maximum Entropy (ME) [7], Naïve Bayes (NB) [7] or Random Forest (RF) [9]. Recurrent neural network (RNNs) with long short-term memory (LSTM) have been used to address long-term dependencies [4, 10], incorporating a coverage mechanism to deal with phrasal ADRs and distinguish ADRs from indications [10]. Variants of convolutional neural networks (CNNs) have also been explored [12]. Finally, graphs have been used to model drug-ADR relationships [9] or interactions between words and candidate phrases [11]. Common errors were due to indications mistaken for ADRs, rare or idiomatic terms, misspellings, ambiguous statements, short posts or generic non-personal statements [2, 3, 4, 7].

In the published literature, ADR detection from social media is approached as a binary classification task, i.e., whether a post contains an ADR, or as an entity recognition and extraction task. Such approaches do not account for the

fact that patients often describe more than one ADR in a single post, for instance, headache and nausea. ADR detection has been formulated as a multi-label learning problem using structured data sources such a PubChem, SIDER or DrugBank [13, 14]. Multi-label classification has also been applied to extract coping strategies following ADRs (e.g., drinking ginger tea after nausea) using data from a Facebook support group [15].

Reddit is a very active social platform with rich textual content accessible through an application programming interface (API). However, compared to Twitter, Reddit has rarely been studied specifically as an ADR detection source [16], although other health topics have been researched [17].

C. Our Approach and Contribution

Using Reddit posts, we have approached the detection of ADR mentions as a multi-label classification problem, which, to the best of our knowledge, has not been performed before. To achieve this, we have created an annotated dataset by supplementing publicly available annotated posts [16] with freshly collected and annotated posts. Unlike Mesbah et al. who split annotated Reddit posts into sentences and whose work focused on detecting the actual ADR span [16], we considered the posts in their entirety, each being associated with a set of labels. The distribution of labels mapped to the Medical Dictionary for Regulatory Activities (MedDRA) has guided the creation of data subsets that varied by the number of labels included within each MedDRA level. On each data subset, we have experimented with different multi-label learning methods associated with different base classifiers.

Our contribution can be summarized in the following points:

- Formulation of the ADR detection in social media posts from Reddit as a multi-label classification task, which, as far as we know, has not been done before.
- Collection and manual annotation of Reddit posts to enrich an existing dataset. This is an effort towards alleviating the problem of the very small number of datasets available for the problem under study. Our dataset is publicly available[1].
- Exploration of the impact of the MedDRA hierarchy on multi-label classification.
- Experimentation with several multi-label problem transformation methods combined with several base classifiers.

The rest of this paper is organised as follows: In section II we provide a brief overview of multi-label classification, with emphasis on the methods and metrics we used in this work. Section III describes how our dataset was created and how labels were streamlined using MedDRA. We describe the experimental setting and discuss our results in sections IV and V, respectively. Finally, in section VI we conclude and outline possible areas for improvement.

II. Multi-label classification

A. Multi-label Problem

A multi-label dataset can be represented with:

- $L = \{\lambda_j : j = 1...q\}$, a finite set of labels of size q,
- $D = \{(x_i, Y_i), i = 1...m\}$, a set of m training examples, where $\mathbf{x_i}$ is the feature vector and $Y_i \subseteq L$ the set of labels of the i-th example [18].

A simple example is shown in Table I.

TABLE I
EXAMPLE OF MULTI-LABEL DATASET [18]

Example	Attributes	Label set
1	$\mathbf{x_1}$	$\{\lambda_1, \lambda_4\}$
2	$\mathbf{x_2}$	$\{\lambda_3, \lambda_4\}$
3	$\mathbf{x_3}$	$\{\lambda_1\}$
4	$\mathbf{x_4}$	$\{\lambda_2, \lambda_3, \lambda_4\}$

B. Multi-Label Learning Algorithms

Multi-Label classification methods have been classified into two categories: *problem transformation* and *algorithm adaptation* [18].

In problem transformation, the learning task is transformed into one or more single-label classification tasks. Binary relevance (BR) is a method that learns q binary classifiers, one for each different label in L [18]. Classifier Chains (CC) is an extension of BR where classifiers are linked along a chain (of length equal to the number of labels q) and the feature space of each link is extended with the label relevance of all previous links [19]. In label powerset (LP) each unique set of labels is considered as one of the classes of a new single-label classification task [18]. The random k-labelsets (RAkEL) method constructs an ensemble of LP classifiers, which are trained using a small random subset of the set of labels [18].

Adapted algorithms extend specific learning algorithms (e.g., k-Nearest Neighbours, decision trees [DT], neural networks) to handle multi-label data directly.

C. Multi-label Dataset Statistics

The performance of a given multi-label learning algorithm may be influenced by the number of labels of each example compared to the size q of the set of labels [18]. The label cardinality of a dataset D is the average number of labels of the examples in D:

$$labelCardinality = \frac{1}{m}\sum_{i=1}^{m}|Y_i|, \tag{1}$$

where m is the number of examples in D and Y_i, the set of labels of the i-th example. The label density of D is the average number of labels of the examples in D divided by q. The number of distinct label sets (classes) is also important for many transformation methods that operate on subsets of labels.

D. Multi-label Classification Metrics

Multi-label classification metrics can be based on instances or on labels. Let us consider an evaluation dataset of multi-label examples $(\mathbf{x_i}, Y_i), i = 1...m$, where $Y_i \subseteq L$ is the set of

[1]https://github.com/unic-ailab/adr_classification

true labels and $L = \{\lambda_j : j = 1...q\}$ is the set of all labels. The set of labels that are predicted by a multi-label learning method for a given instance $\mathbf{x_i}$, is denoted as Z_i.

Example-based measures consider the average differences of the actual and the predicted sets of labels over all examples of the evaluation dataset. Among these measures, the Hamming loss is the fraction of labels that are incorrectly classified:

$$HammingLoss = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \Delta Z_i|}{q}, \quad (2)$$

where Δ is the symmetric difference of the two sets, m the number of examples and q the number of labels [18].

F_1 is the harmonic mean of precision and recall [20].

Label-based methods separate evaluations for each label, which are subsequently averaged over all labels. Label-based measures can be based on any binary evaluation measure, followed by an averaging operation (macro- or micro-) for all labels [18].

III. Data collection, wrangling and analysis

A. Dataset Creation

Reddit[2] is an online discussion and community platform founded in 2005. It is formed of communities ('subreddits') created and moderated by users ('redditors'). Subreddit members abide by community rules and can post links, images, videos or text. With over 57 million daily users, 100,000 subreddits and 13 billion posts and comments, Reddit is one of the most popular social platforms on the web. Reddit posts, comments, and metadata can be accessed via the site itself, or via a publicly available API. Numerous subreddits are dedicated to health-related topics and address medical conditions and their management, in a variety of contexts such as personal testimonies, exchange of advice, discussion on research, etc [17].

Mesbah et al. [16] collected 1,626 Reddit posts containing at least one of the drug names used in previous research [3] and recruited a medical expert to annotate them for ADR mentions. Of the annotated posts, 599 contained at least one ADR. The corresponding post identifiers and ADR annotations were made publicly available by the authors as 'ADR_EMNLP2019'[3]. Using the Python Reddit API Wrapper (PRAW)[4], we attempted to retrieve from Reddit more information about the corresponding posts including subreddit, title and text. Only 318 posts (53%) were retrieved with all the requested information. Ten (10) posts were restricted from access, 271 posts could be retrieved but their text had been deleted. To complement the 318 annotated posts available from 'ADR_EMNLP2019', we scraped Reddit for posts mentioning one or more of the drug names referred to earlier [3]. A total of 586 posts were retrieved, and manually examined for ADR mentions by one of the authors, a pharmacovigilance expert by background. Of the freshly collected posts, 219 mentioned at least one ADR, which added to the posts from 'ADR_EMNLP2019', resulted in a dataset of 537 annotated posts (see Fig. 1).

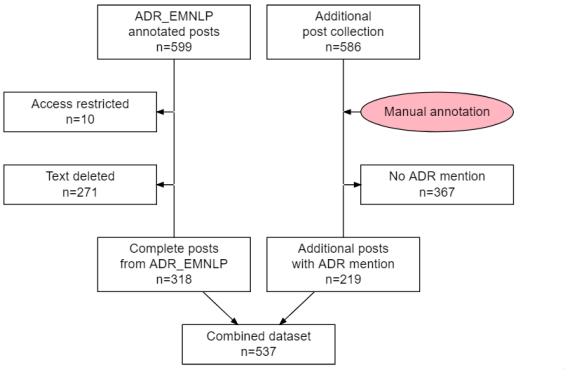

Fig. 1. Data collection and annotation overview

When considering the annotations 'as-is', the combined dataset was associated with a large number of different labels (n=755), greater than the number of posts. There was indeed an 'artificial' granularity as the exact same events were tagged in multiple ways (e.g., agitated/agitation, dizziness/dizzy/dizzy spells), in addition to misspellings. The next step was therefore to map all labels against standardised medical concepts, and for this the MedDRA dictionary was used.

B. Label Normalisation

MedDRA[5] was developed in the late 1990s by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) as a standardised medical terminology to facilitate sharing of regulatory information for medicinal products used by humans . There are five levels to the MedDRA hierarchy, arranged from specific to general: Low Level Term (LLT, n=86,714, e.g., 'Feeling queasy'), Preferred Term (PT, n=25,916, e.g., 'Nausea'), High Level Term (HLT, n=1,737, e.g., 'Nausea and vomiting symptoms'), High Level Group Term (HLGT, n=337, e.g., 'Gastrointestinal signs and symptoms'), system organ class (SOC, n=27, e.g., 'Gastrointestinal disorders')[6].

Each label of the combined dataset was manually assigned to a MedDRA PT, which in turn was mapped to higher MedDRA levels. For instance, the original labels 'dizzy', 'dizziness' and 'dizzy spells' were all mapped to a single PT: 'Dizziness'. The operation greatly reduced the overall number of labels compared with the 755 original annotations: 363 PT labels (2-fold reduction), 193 HLT labels (4-fold reduction), 106 HLGT labels (7-fold reduction), 24 SOC labels (31-fold reduction).

C. Label Exploratory Analysis

Regardless of the MedDRA level, most posts had between one (peak of the distribution) and 3 labels, but there were

[2]https://www.redditinc.com/

[3]https://github.com/mesbahs/ADR_EMNLP

[4]https://praw.readthedocs.io/en/stable/

[5]https://www.meddra.org/

[6]Numbers of terms are for MedDRA version 26.0

outliers with e.g., more than 30 labels at PT level. The most frequent labels were 'Psychiatric disorders' at SOC level (n=282), 'General system disorders NEC[7]' at HLGT level (n=121), 'Withdrawal and rebound effects' (n=76) at HLT level, and 'Withdrawal syndrome' at PT level (n=76). The imbalance in labels appeared to increase when moving up the MedDRA hierarchy, i.e., it was the most pronounced at SOC level.

To further streamline the set of labels, the impact of label selection on the dataset size was explored. Within each MedDRA level, each term was ranked from the most to the least frequent, and the cumulative frequency of occurrence was calculated. The idea was to find an acceptable trade-off to reduce the set of labels without compromising the size of the dataset too much. For example, by keeping only the 20 most frequent HLGTs, 465 posts, i.e., 87% of the dataset would be retained. This is illustrated in Fig. 2 and this has guided the experimental set-up described in the next section.

IV. Experimentation

A. Data Sub-setting

Twelve (12) variants of the dataset were prepared using the 4 different MedDRA levels and, within each MedDRA level, 3 selections were applied based on rank: i) top x labels, ii) top y labels, with x and y being numbers chosen arbitrarily based on our coverage analysis, and iii) all labels, i.e., the full dataset annotated at that MedDRA level. Each dataset was named after the MedDRA level and the number of labels included, e.g., 'PT-20'. For each dataset, the number of posts, the number of distinct labels (which may be higher than the corresponding rank in case of 'ties'), the number of distinct sets of labels, the cardinality and the density were calculated. These are shown in Table II.

TABLE II
Characteristics of datasets

dataset	posts	features	labels	sets	cardinality	density
PT-20	361	4,825	21	137	1.712	0.086
PT-50	438	5,353	57	260	2.123	0.042
PT-ALL	537	6,029	363	424	2.739	0.005
HLT-20	416	5,347	20	172	1.858	0.093
HLT-40	460	5,541	42	261	2.241	0.056
HLT-ALL	537	6,029	193	381	2.605	0.005
HLGT-20	465	5,636	20	199	2.082	0.104
HLGT-30	494	5,727	31	242	2.225	0.074
HLGT-ALL	537	6,029	106	317	2.421	0.005
SOC-07	500	5,901	7	61	1.720	0.246
SOC-12	521	5,974	12	110	1.839	0.153
SOC-ALL	537	6,029	24	144	1.909	0.004

B. Pre-processing

For the whole corpus, the title and the text body of each post were concatenated, as the former was also taken into account consideration during annotation. The resulting text was pre-processed: the case was normalised, stop-words, numbers and special characters were removed, the text was tokenized and the words were stemmed. For each dataset, labels were hot-encoded.

Feature extraction was performed using Term Frequency - Inverse Document Frequency (TF-IDF), resulting in approximately 5,000 to 6,000 features in each dataset (see Table II).

C. Models

Each dataset was split into a training (70%) and test (30%) set. The four multi-label learning methods introduced in Section II-B, BR, CC, LP and RAkEL, were applied to each training set. For each method, 4 base classifiers were tested: DT, multinomial NB, RF and SVM. Scikit-learn[8] and scikit-multilearn[9] were used for the base classifiers and the multi-label methods, respectively [21, 22]. A linear kernel was applied for SVM, a random seed was used for DT, RF and SVM for reproducibility, and the base classifier requiring matrices [input, output] in dense representation was set for RAkEL. All other parameters were left as default.

V. Evaluation

A. Results

The performance was evaluated on each test set using the micro F1 score (Fig. 3), macro F1 score (data not shown) and Hamming loss (Fig. 4).

The best F1-scores micro and macro were obtained with CC-DT on PT-20 (61.4% and 56.0%, respectively). The best hamming loss (0.008) was obtained with BR-NB, BR-RF, CC-NB, CC-RF, RAkEL-NB and RAkEL-RF, all on PT-ALL.

1) Impact of dataset: F-scores were generally better on the SOC datasets, particularly SOC-07. Within each MedDRA level, it appears that the higher the number of labels, the lower the score. On the contrary, the hamming loss appeared better at PT level. It also appears that datasets had a stronger influence on the hamming loss than classification algorithms.

2) Impact of base classifier: DT was the best classifier in many experiments in terms of F-scores. NB was only able to make correct predictions on the SOC-07 and/or in association with LP; in all other settings, F-scores were nil.

3) Impact of multi-label learning algorithm: The best F-scores were obtained with BR and CC, but interestingly, in association with NB or RF at the PT, HLT and HLGT levels, LP did give slightly better results than BR/CC.

B. Error Analysis

A random sample of 10 predictions made by two models on two different datasets, BR-DT on PT-20 and RAkEL-DT on SOC-ALL, respectively, were examined to gain more insights on the weaknesses and strengths of the models. Details of 3 erroneous predictions by each model can be seen in Fig. 5.

In the first example, the model correctly predicted 'Suicidal ideation' but wrongly detected 'Anxiety' (false positive), which in that context was a therapeutic indication, not an

[7]NEC: not elsewhere classified, denotes groupings of miscellaneous terms that do not readily fit into other hierarchical classifications

[8]https://scikit-learn.org/stable/index.html

[9]http://scikit.ml/index.html

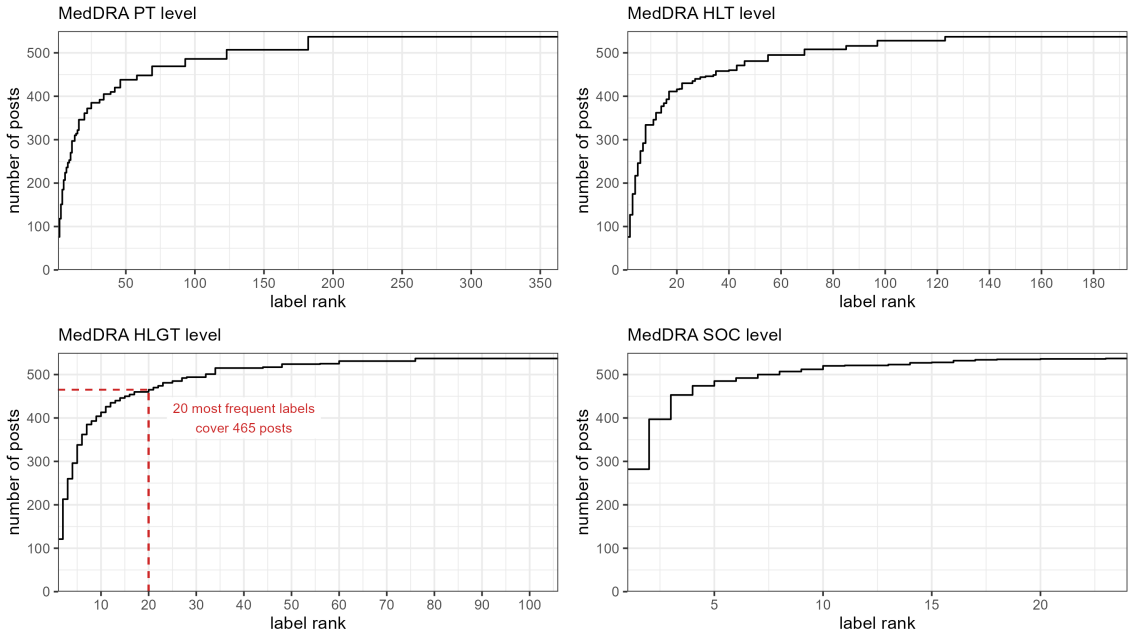

Fig. 2. Dataset coverage depending on number of labels retained

model	PT-20	PT-50	PT-ALL	HLT-20	HLT-40	HLT-ALL	HLGT-20	HLGT-30	HLGT-ALL	SOC-07	SOC-12	SOC-ALL
BR-DT	0.612	0.558	0.352	0.561	0.480	0.393	0.498	0.487	0.416	0.609	0.551	0.528
BR-NB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.533	0.398	0.307
BR-RF	0.000	0.007	0.000	0.033	0.007	0.009	0.058	0.046	0.029	0.541	0.449	0.428
BR-SVM	0.127	0.101	nan	0.117	0.165	nan	0.202	0.219	nan	0.564	0.515	nan
CC-DT	0.614	0.570	0.362	0.568	0.487	0.391	0.499	0.484	0.413	0.591	0.542	0.507
CC-NB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.525	0.399	0.307
CC-RF	0.021	0.013	0.004	0.017	0.015	0.009	0.026	0.035	0.025	0.549	0.448	0.405
CC-SVM	0.118	0.101	nan	0.124	0.159	nan	0.211	0.224	nan	0.567	0.514	nan
LP-DT	0.279	0.162	0.107	0.371	0.234	0.139	0.262	0.255	0.190	0.456	0.440	0.363
LP-NB	0.147	0.139	0.078	0.088	0.108	0.054	0.132	0.094	0.067	0.496	0.382	0.314
LP-RF	0.312	0.249	0.165	0.266	0.251	0.189	0.286	0.297	0.187	0.508	0.430	0.347
LP-SVM	0.332	0.236	nan	0.296	0.265	nan	0.341	0.289	nan	0.539	0.514	nan
RK-DT	0.578	0.544	0.368	0.528	0.454	0.382	0.465	0.478	0.419	0.547	0.524	0.508
RK-NB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.496	0.394	0.306
RK-RF	0.000	0.033	0.009	0.065	0.015	0.018	0.094	0.073	0.048	0.562	0.439	0.413
RK-SVM	0.136	0.107	nan	0.132	0.164	nan	0.197	0.249	nan	0.545	0.493	nan

Fig. 3. F1-score (micro)
BR: binary relevance, CC: classifier chains, DT: decision tree, LP: label powerset, NB: Naïve Bayes, RF: Random Forest, RK: random k-labelsets (RAkEL), SVM: support vector machine.

model	PT-20	PT-50	PT-ALL	HLT-20	HLT-40	HLT-ALL	HLGT-20	HLGT-30	HLGT-ALL	SOC-07	SOC-12	SOC-ALL
BR-DT	0.070	0.033	0.010	0.089	0.052	0.016	0.110	0.073	0.026	0.180	0.137	0.075
BR-NB	0.083	0.040	0.008	0.096	0.047	0.014	0.107	0.074	0.023	0.175	0.136	0.083
BR-RF	0.083	0.040	0.008	0.094	0.046	0.014	0.105	0.072	0.023	0.167	0.118	0.065
BR-SVM	0.078	0.038	nan	0.091	0.044	nan	0.099	0.066	nan	0.168	0.113	nan
CC-DT	0.071	0.032	0.010	0.083	0.051	0.016	0.107	0.073	0.026	0.186	0.134	0.076
CC-NB	0.083	0.040	0.008	0.096	0.047	0.014	0.107	0.074	0.023	0.179	0.136	0.083
CC-RF	0.083	0.040	0.008	0.095	0.046	0.014	0.107	0.072	0.023	0.161	0.119	0.068
CC-SVM	0.079	0.038	nan	0.091	0.044	nan	0.099	0.066	nan	0.170	0.115	nan
LP-DT	0.115	0.066	0.012	0.110	0.076	0.022	0.145	0.114	0.037	0.246	0.161	0.095
LP-NB	0.112	0.049	0.010	0.133	0.063	0.018	0.137	0.096	0.031	0.190	0.141	0.084
LP-RF	0.090	0.044	0.010	0.108	0.053	0.016	0.114	0.075	0.027	0.187	0.131	0.081
LP-SVM	0.095	0.052	nan	0.112	0.060	nan	0.115	0.084	nan	0.186	0.120	nan
RK-DT	0.068	0.034	0.009	0.090	0.055	0.016	0.113	0.076	0.026	0.204	0.135	0.075
RK-NB	0.083	0.040	0.008	0.096	0.047	0.014	0.107	0.074	0.023	0.190	0.135	0.083
RK-RF	0.083	0.039	0.008	0.093	0.046	0.014	0.103	0.071	0.023	0.160	0.125	0.068
RK-SVM	0.078	0.038	nan	0.090	0.044	nan	0.099	0.065	nan	0.178	0.119	nan

Fig. 4. Hamming loss
BR: binary relevance, CC: classifier chains, DT: decision tree, LP: label powerset, NB: Naïve Bayes, RF: Random Forest, RK: random k-labelsets (RAkEL), SVM: support vector machine.

ADR. In the second example, the model's predictions, although not matching the annotation, could actually be considered more complete: 'Drug tolerance' and 'Feeling abnormal' were in fact described in the post, and 'Drug dependence' and 'Dependence' may be considered equivalent in this context. In the third example, the model correctly identified 'Headache', but in addition predicted 'Irritability' (false positive) likely based on the mention of eye irritation in the post, which is a different medical issue. In the fourth and fifth examples (false negatives), several references to sexual and mood issues (SOC 'Psychiatric disorders'), and shortness of breath, respectively, were not detected by the model. In the sixth example, the model detected a nervous disorder (false positive), but it is unclear from the post what may have led to this prediction. Amongst the other errors, not shown on the figure, there were more false positives (e.g., predicted terms mentioned in the text but not as ADRs), false negatives with ADRs not detected by the model with no obvious reason (e.g., nausea) and in one instance the model predicted 'Weight increased' whereas the redditor had in fact dropped weight.

VI. Discussion

Our best F-scores are in line with published results using multi-label text datasets ([15, 19, 23]), although these were meant to address different problems and the comparison may

Original text abstracts	Experiment	True labels	Predicted labels
[...] She has been diagnosed with generalized anxiety syndrome [...] Doctors just give her anti anxiety [...] has suicidal thoughts now [...].	BR-DT on PT-20	Suicidal ideation	Suicidal ideation, Anxiety
[...] im addicted to lyrica and baclofen [...] the tolerance increased [...] feel slight buzz [...]	BR-DT on PT-20	Drug dependence	Drug tolerance, Feeling abnormal, Dependence
[...] been getting bad headaches [...] my eyes feel quite irritated.	BR-DT on PT-20	Headache	Irritability, Headache
[...] I have no interest in sex, my emotions are disappearing, [...] not happy either. [...] nauseous all the time [...] losing my sex drive [...] I have even less motivation to do things [...]	RK-DT on SOC-ALL	Gastrointestinal, Psychiatric	Gastrointestinal
[...] I feel tired all the time and short of breath. [...]	RK-DT on SOC-ALL	Respiratory, General	General
[...] struggling with extreme irritability lately. [...] terrible irritability. [...] just raged out [...] anger issues [...]	RK-DT on SOC-ALL	Psychiatric	Nervous, Psychiatric

Fig. 5. Analysis of a sample of erroneous predictions

not be relevant. Some of these datasets had similar label cardinality and density to ours, but one notable difference is the high number of features for a relatively small dataset in our case.

Yapp et al. [19] evaluated several base classifiers accross several benchmark datasets and found that the best classifier depends on the multi-label learner, e.g., SVM for BR, CC, RAkEL, DT for ensembles of classifier chains. In contrast, DT outperformed SVM in most of our settings.

The best F-scores were obtained on the two datasets (SOC-07 and PT-20) with the lowest cardinalities (1.720 and 1.712, respectively). Bernardini et al. [24] observed a high correlation between cardinality and density and the performance of several multi-label learners on a music dataset. A possible additional explanation for the generally better performance of the models at PT and SOC levels, compared to HLT and HLGT, could also be related to semantic and training aspects: on the one hand, PTs, being the most specific terms, may be closer to the verbatim text (e.g. 'nausea'); on the other hand, SOCs, due to their lower number, may be associated with more training examples. Our results suggest that further research in a similar setting could focus on these two MedDRA levels to reduce the number of dataset variants.

The error analysis (see section V-B) suggests that contextual information may not be sufficiently addressed in our models. For instance, the expression 'dropped my weight' led to a post being wrongly classified as 'Weight increased'.

Finally, there may have been differences in annotation approaches between ADR_EMNLP [16] and our fresh collection of posts, which could have an impact on training and performance.

VII. Conclusion and Future Work

In this article, we approached the detection of mentions of ADRs in social media posts from the platform Reddit as a multi-label classification problem and explored the impact of the MedDRA hierarchy, which, to the best of our knowledge, has not been studied before.

Our results are promising considering the small size of the dataset and the high numbers of label sets and features. Several actions may improve the performance and generalization of our models: (i) The size of the training set should be increased by collecting and annotating more posts. (ii) A dimensionality reduction technique (feature selection or extraction) should be applied to address the high number of features. (iii) Other word embedding approaches, and techniques that help account for the context surrounding the ADR mentions should be explored. (iv) The hyperparameters of base classifiers (DT, NB) should be tuned. (v) Other multi-label learning methods and/or base classifiers should be explored. These include adapted algorithms such as multilabel k Nearest Neighbours (MLkNN) and deep learning approaches. If sufficient performance is reached after these actions, a system could be implemented whereby new posts are streamed from Reddit and ADRs or sets of ADRs, if any, are detected and normalised with a view to support the analysis and monitoring of reporting trends.

Acknowledgment

The first author would like to thank Izabela Skibicka-Stępień and Giulia Gabrielli for their review of the manuscript.

Disclaimer

The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the organisations with which the authors are affiliated.

References

[1] WHO. *The importance of pharmacovigilance.* World Health Organization, 2002.

[2] R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. "Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks". In: *Proceedings of the 2010 workshop on biomedical natural language processing*. 2010, pp. 117–125.

[3] A. Nikfarjam, A. Sarker, K. O'connor, R. Ginn, and G. Gonzalez. "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features". In: *Journal of the American Medical Informatics Association* 22.3 (2015), pp. 671–681.

[4] A. Cocos, A. G. Fiks, and A. J. Masino. "Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts". In: *Journal of the American Medical Informatics Association* 24.4 (2017), pp. 813–821.

[5] S. Comfort et al. "Sorting through the safety data haystack: using machine learning to identify individual case safety reports in social-digital media". In: *Drug safety* 41 (2018), pp. 579–590.

[6] A. Magge et al. "DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter". In: *Journal of the American Medical Informatics Association* 28.10 (2021), pp. 2184–2192.

[7] A. Sarker and G. Gonzalez. "Portable automatic text classification for adverse drug reaction detection via multi-corpus training". In: *Journal of biomedical informatics* 53 (2015), pp. 196–207.

[8] D. Bollegala, S. Maskell, R. Sloane, J. Hajne, and M. Pirmohamed. "Causality patterns for detecting adverse drug reactions from social media: text mining approach". In: *JMIR public health and surveillance* 4.2 (2018), e8214.

[9] R. Eshleman and R. Singh. "Leveraging graph topology and semantic context for pharmacovigilance through twitter-streams". In: *BMC bioinformatics*. Vol. 17. 13. BioMed Central. 2016, pp. 77–93.

[10] S. Chowdhury, C. Zhang, and P. S. Yu. "Multi-task pharmacovigilance mining from social media posts". In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 117–126.

[11] Z. Li, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang. "Lexicon knowledge boosted interaction graph network for adverse drug reaction recognition from social media". In: *IEEE Journal of Biomedical and Health Informatics* 25.7 (2020), pp. 2777–2786.

[12] M. Rakhsha, M. R. Keyvanpour, and S. V. Shojaedini. "Detecting adverse drug reactions from social media based on multichannel convolutional neural networks modified by support vector machine". In: *2021 7th International Conference on Web Research (ICWR)*. IEEE. 2021, pp. 48–52.

[13] E. Muñoz, V. Nováček, and P.-Y. Vandenbussche. "Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models". In: *Briefings in bioinformatics* 20.1 (2019), pp. 190–202.

[14] D. Afdhal, K. W. Ananta, and W. S. Hartono. "Adverse drug reactions prediction using multi-label linear discriminant analysis and multi-label learning". In: *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE. 2020, pp. 69–76.

[15] A. Dirkson, S. Verberne, G. van Oortmerssen, H. Gelderblom, and W. Kraaij. "How do others cope? Extracting coping strategies for adverse drug events from social media". In: *Journal of Biomedical Informatics* 139 (2023), p. 104228.

[16] S. Mesbah et al. "Training data augmentation for detecting adverse drug reactions in user-generated content". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 2349–2359.

[17] U. Naseem, M. Khushi, J. Kim, and A. G. Dunn. "RHMD: a real-world dataset for health mention classification on Reddit". In: *IEEE Transactions on Computational Social Systems* (2022).

[18] G. Tsoumakas, I. Katakis, and I. Vlahavas. "Mining multi-label data". In: *Data mining and knowledge discovery handbook* (2010), pp. 667–685.

[19] E. K. Yapp, X. Li, W. F. Lu, and P. S. Tan. "Comparison of base classifiers for multi-label learning". In: *Neurocomputing* 394 (2020), pp. 51–60.

[20] S. Godbole and S. Sarawagi. "Discriminative methods for multi-labeled classification". In: *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings 8*. Springer. 2004, pp. 22–30.

[21] F. Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[22] P. Szymański and T. Kajdanowicz. "A scikit-based Python environment for performing multi-label classification". In: *arXiv preprint arXiv:1702.01460* (2017).

[23] R. Venkatesan and M. J. Er. "Multi-label classification method based on extreme learning machines". In: *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*. IEEE. 2014, pp. 619–624.

[24] F. C. Bernardini, R. B. Da Silva, E. Meza, and R. das Ostras-RJ-Brazil. "Analyzing the influence of cardinality and density characteristics on multi-label learning". In: *Proc. X Encontro Nacional de Inteligencia Artificial e Computacional-ENIAC* (2013).