Hierarchical Classification of Adverse Events Based on Consumer's Comments

Monika Kaczorowska[1,2][0000-0002-4618-7937], Piotr Szymczak[1] and Sergiy Tkachuk[1,3][0000-0002-3434-6320]

[1] Reckitt Benckiser Group, Global Data & Analytics, Zajęcza 15, 00-351, Warsaw, Poland
[2]Polish-Japanese Academyof Information Technology, Koszykowa 86, 02-008, Warsaw, Poland
[3]Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447, Warsaw, Poland
monika.kaczorowska@reckitt.com,
piotrwojciech.szymczak@reckitt.com, sergiy.tkachuk@reckitt.com

Abstract. This paper focuses on autonomously classifying adverse events based on consumers' comments regarding health and hygiene products. The data, comprising over 152,000 comments, were collected from e-commerce sources and social media. In the present research, we propose a language-independent approach using machine translation, allowing for unified analysis of data from various countries. Furthermore, this study presents a real-life application, making it potentially beneficial for subsequent scientific research and other business applications. A distinguishing feature of our approach is the efficient modeling of colloquial language instead of medical jargon, which is often the focus of adverse event research. Both hierarchical and non-hierarchical classification approaches were tested using Random Forest and XGBoost classifiers. The proposed feature extraction and selection process enabled us to include tokens important to minority classes in the dictionary. The F1 score was utilized to quantitatively assess the quality of classification. Hierarchical classification allowed for faster classification processes than the non-hierarchical approach for the XGBoost classifier. We obtained promising results for XGBoost; however, further research on a wider range of categories is required.
Keywords: classification, adverse event, NLP.

1 Introduction

Adverse events are defined as untoward medical occurrences following exposure to a medicine, not necessarily caused by that medicine [1]. Adverse events can potentially be hazardous to humans, causing irreversible changes in the human body and, in extreme cases, death [2]. They pose a significant public issue, affecting human health and life, and causing substantial financial losses [3, 4]. Adverse events can be caused by pharmaceutical products, cosmetics, care products, and cleaning agents [5]. The Center for Food Safety and Applied Nutrition (CFSAN) has provided the Adverse Event Reporting System (AERS), which allows for reporting adverse events and

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17

product complaints related to foods, dietary supplements, and cosmetics [6]. Knowledge of adverse event occurrences enables improvements in products, making them safer and more attractive to consumers.

Nowadays, machine learning techniques are applied in various fields, including adverse event classification [7, 8]. Common approaches are based on using Support Vector Machines [4], Random Forest [7], or Maximum Entropy [9]. Neural networks, such as Convolutional Neural Networks (CNN) [10], attention-based deep neural networks [11], and Bidirectional Encoder Representations from Transformers (BERT) [12], are also employed in adverse event classification. It is worth noting that imbalanced classes are a common problem in this area [13].

Recently, the role of social media and e-commerce portals has become increasingly crucial for expressing opinions among consumers. Social media [14] and e-commerce data sources pose challenges from a data processing perspective, especially for Natural Language Processing. Comments and opinions published on the Internet often contain misspellings and slang expressions. However, activity on social media or e-commerce platforms is an essential part of our lives [15] and can be considered a valuable and underexplored source of information about adverse events [13]. Pattanayak et al. discussed the advantages of applying e-commerce insights in the pharmaceutical industry [16].

Our literature review reveals that approaches to adverse event classification primarily focus on binary classification and are related to adverse events occurring in drugs. Additionally, the conducted analyses concentrate on texts initially written in one language, usually English. Language independence allows for broader analysis of adverse events, taking cultural trends into consideration and enabling quicker detection of problems. In the reviewed scientific papers, authors apply classical classifiers and neural networks; however, to the best of our knowledge, the hierarchical classification approach has not been used for the adverse event classification problem. Hierarchical classification is described as dividing a problem into smaller classification problems [17, 18].

In this paper, we propose the first attempt to conduct hierarchical classification on a large dataset consisting of e-commerce and social media texts. Our research focuses on over-the-counter (OTC) drugs and other health and hygiene products, including sexual well-being products and household chemicals. Our main contributions within the framework of the presented research are:

- performing multiclass classification of adverse events based on consumers' comments using a language-agnostic approach,
- examining both hierarchical and non-hierarchical approaches to classification,
- carrying out classification in conditions as close to production as possible,
- expanding research on adverse event classification to non-drug products.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17

2 Materials and methods

2.1 Dataset

The dataset consists of over 152,000 texts gathered from e-commerce sources such as Amazon and Lazada, and social media platforms like Facebook and Twitter. It includes, among others, online product reviews, discussions about usage, and messages directed to brand profiles. This proprietary dataset was collected with the assistance of several third parties for internal processes related to customer relationship management and adverse event reporting. The texts were written in English and other languages, such as Spanish, Russian, Japanese, and Arabic. Approximately 70% of the data were initially written in English, while the remaining 30% were written in non-English languages and machine-translated into English. It is worth noting that the dataset reflects the actual distribution of the data collected, with heavily imbalanced classes – see Fig.1. The texts were used as collected from the Internet and are unstructured, containing grammar mistakes, misspellings, numbers, and emojis. Non-English observations were machine-translated into English, which can make classification more challenging by introducing translation artifacts and mistakes. However, this approach allows for language-independent classification. Customer relations agents labeled the data in accordance with internal procedures for handling online engagements from customers, ensuring high data quality as trained subject matter experts processed it. A hierarchical labeling system was applied; for example, each observation was assigned a label on each hierarchy level. Fig.1 shows the category tree structure, with the percentage of observations of each class in the adverse-event and non-adverse event sets given in parentheses. The label "Other" means that the observation contains adverse events other than the ones listed.

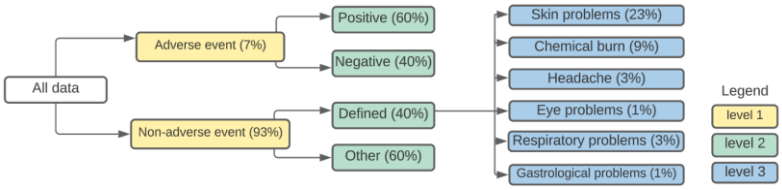

Fig. 1. Structure of labelled data

2.2 Data Processing

The data processing procedure included data cleaning, data translation, feature extraction and selection, and feature normalization. Data cleaning consisted of removing emojis and links. The machine translation process was applied to analyze all gathered data, even if it was not initially written in English. English has been defined as a common language for all comments. Text Translator, using Azure Cognitive Services [19, 20], was applied to translate non-English comments. It is worth noting that machine translation is not a perfect mechanism, and the translation of concise phrases can result in over-translation [21]. To assess the quality of translation, the following pilot experiment was conducted: English language comments were used as a training

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17

dataset. The tests were performed separately on English and translated comments. The tests were repeated, and the obtained results showed that classification metrics did not differ, hence the quality of translation was sufficient.

Fig. 2 shows the feature extraction flowchart. The extracted features can be divided into two groups. The first group contains a set of general features, calculated once before the classification process. This set includes the number of words in each comment, the number of sentences, the ratio of stop words to all tokens, the number of exclamation signs, the number of questions, the ratio of uppercase to all tokens, mean word length, mean sentence length, and the ratio of numbers to all tokens. The second set of features was calculated before each level of classification. The N most common tokens were extracted for each class, with the union of these extracts being used as a dictionary for vectorization. This approach enables taking into account tokens from all classes. Otherwise, tokens describing small-sized classes could be omitted. In the present research, the first 500 tokens for each class were extracted and vectorized using the Term Frequency-Inverse Document Frequency method (TF-IDF). The Min-Max Scaler, which scales the minimum and maximum values to 0 and 1, was applied. We implemented the software for data analysis and classification in Python, using libraries such as NumPy, scikit-learn, pandas, and Azure-specific libraries.

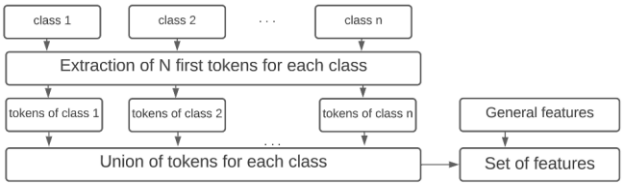

Fig. 2. Feature extraction flowchart.

2.3 Classification

The main aim of the classification process was to predict if the comment mentions an adverse event and, if yes, to assign the type of adverse event. The additional objective was to predict the sentiment of non-adverse events. Two tree structure classification methods were tested: XGBoost and Random Forest. Class weights were provided in the binary classification problem on the first level (i.e., adverse event vs. not adverse event classification) to tackle the class imbalance problem. The rest of the parameters were set empirically. Parameter tuning was conducted using grid search, and the following parameters were tested: learning rate and maximum depth of tree for XGBoost, and tree number for Random Forest. The following parameters were set for XGBoost [22]: learning rate: 0.3, maximum depth of a tree: 6, and scale_pos_weight was used. For Random Forest, the chosen parameters were [23]: tree number: 100, class_weight: 'balanced'. These classification models were chosen for the following reasons: a lower number of parameters (compared to deep learning models), lower computational cost allowing experiments to run in a reasonable amount of time, and clear interpretability. Moreover, literature shows that these classifiers can be applied to text classification problems [24, 25] with imbalanced classes successfully.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17

Two approaches were tested: the proposed hierarchical approach and the non-hierarchical approach. The hierarchical approach is presented in Fig. 3. In the non-hierarchical approach, all nine classes (tree leaves): seven classes of adverse events and two classes of non-adverse events were considered, and 9-class classification was performed. Hierarchical classification allows performing classification at different levels of the hierarchy tree sequentially. Separate classifiers are trained using 5-fold cross-validation to better estimate the classifier's power. Various classification thresholds are tested to obtain the best threshold, maximizing the F1 score. The dataset was shuffled and divided into train and test datasets in a stratified way, using an 80:20 proportion. The classifiers were evaluated using the following metrics: accuracy, recall, precision, and F1.

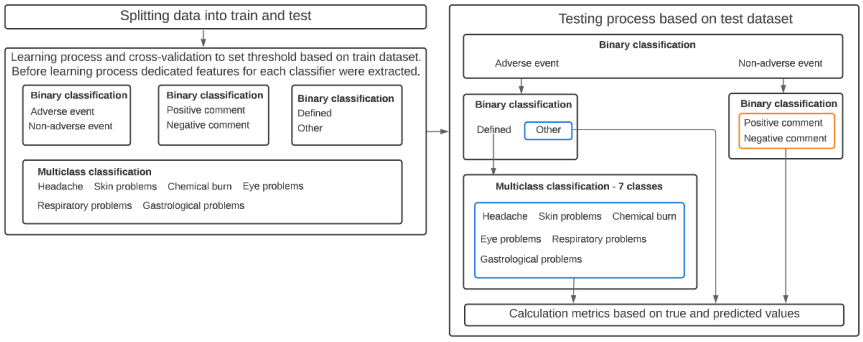

Fig. 3. Hierarchical approach

3 Results

Table 1 presents the results of the classification for hierarchical and non-hierarchical approaches. The table includes the results of nine-class classifications containing seven adverse event classes (six defined classes + seventh – other) and two non-adverse event classes (positive and negative comments). We do not report overall scores because of the class imbalance. The metrics were calculated separately for each class and presented in the table below. We evaluated the model both on training and testing datasets to check if the model is not overfitting. We report only metrics for the test dataset. Precision, recall, and F1 metrics are reported for each class. For adverse events detection, it's crucial to detect as many occurring cases as possible. Because of that, the best metrics for the task are recall and F1 score. The F1 metric was presented for both classifiers. The second column in the table contains information if the class is an adverse event (ADR) or not (NONADR) for top-level binary classification.

The application of hierarchical classification allows adjusting classification thresholds on binary levels in the hierarchy. For XGBoost, the thresholds were set to 0.5 for both binary levels, while for the Random Forest, the threshold was set to 0.3 for the level classifying the occurrence of an adverse event and 0.5 for the level distinguishing between a non-defined adverse event (Other) and the defined types of adverse events. These thresholds were chosen in the course of preliminary tests. As it may be

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17

noticed, better results were obtained for the XGBoost classifier, especially for minority classes such as gastrological and respiratory problems. The hierarchical approach has given better results for XGBoost for some classes, such as eye problems, gastrological problems, headaches, and skin problems. Using Random Forest, better results were achieved using the hierarchical approach for all defined adverse event classes.

Table 1. Results of non-hierarchical and hierarchical classification for F1measure

Class	**Type**	***Non-hierarchical***		***Hierarchical***	
		XGBoost	Random Forest	XGBoost	Random Forest
Respiratory problems	ADR	0.54	0.06	0.44	0.35
Chemical burn	ADR	0.62	0.49	0.61	0.52
Eye problems	ADR	0.32	0.22	0.56	0.52
Gastrological problems	ADR	0.14	0.01	0.24	0.08
Headache	ADR	0.55	0.21	0.64	0.62
Skin problems	ADR	0.31	0.47	0.52	0.56
Other	ADR	0.67	0.61	0.63	0.68
Positive comment	NONADR	0.88	0.88	0.87	0.88
Negative comment	NONADR	0.81	0.81	0.79	0.80

Additionally, computation time for both approaches was measured. All calculations were carried out in the same conditions regarding hardware, train, and test datasets. One cycle of the feature extraction and the learning process was considered. Performing hierarchical classification using the XGBoost classifier is over 4.5 times faster than a non-hierarchical approach (around 3 minutes for hierarchical and 14 minutes for non-hierarchical). These gains may not seem high, but they will scale with the size of the dataset, the number of classes, and in exhaustive grid searches. In the case of a parameter grid containing a total of 50 combinations, the time gain would rise to 500 minutes. In the case of Random Forest, the hierarchical approach took around 6 minutes, while the non-hierarchical approach took 4 minutes.

4 Discussion and conclusions

The study aimed to classify adverse events independently of comment language using a hierarchical classification based. The classification was performed based on consumers' comments gathered from e-commerce and social-media sources. Our research presents the classification of adverse events occurring for OTC drugs and other medicinal and hygiene products. The goal of our work was not only to obtain better results of adverse events detection but also to create the framework allowing for easier development of a classification model with other levels and classes. The examination of the language-independent approach has shown that it is possible to analyze comments in various languages by mapping them to English using a machine translation mechanism. To the best of our knowledge, it is the first study performed on such an

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17

amount of data independently from comment language, applying hierarchical classification in adverse event problems.

The results showed that the hierarchical approach allows not only to obtain better or similar results but also to carry out calculations in a notably shorter time – over than 4,5 times for XGBoost. Additionally, the hierarchical approach enables threshold adjustment at each binary level, and this approach can be easily extended to more levels. Finding the proper threshold allows us to obtain better results than using the default ones. Hierarchical classification enables the addition of additional levels and classes without starting the learning process from scratch. The application of interpretable machine learning models such as XGBoost and Random Forest allows to create the ranking of features and perform the linguistic analysis of them. Additionally, the hierarchical approach allows for classification with more classes. The application of feature selection technique based on selecting N first tokens for each class allowed us to better tackle class imbalance. Another valuable aspect of our work is the fact that the presented study reflects the real business problem with imbalanced classes. Not all possible categories had been used in the study (class Other) due to insufficient data. This is an area worth exploring in further research. Our proposed method is less time-consuming and gives promising results.

The presented research opens a series of planned studies. We would like to test the approach using more classification algorithms, such as Logistic Regression, Support Vector Machine and Neural Networks. Our present study does not consider the case of multilabel classification (more than one class possible for each of the texts). Further research will also include more detailed class hierarchy levels. We also plan to develop a named entity recognition (NER) model to recognize the drugs (products), indications and individual adverse events as they are reported in social media and e-commerce texts, to support pharmacovigilance practice with detection in search of early product issue signals from the Internet. We can also look to extend the interpretability of the trained models using SHAPley values.

References

1. https://www.ema.europa.eu/en/glossary/adverse-event. Accessed January 27 (2023).
2. Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., ..., Breckenridge, A. M.: Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. Bmj, 329(7456), pp. 15-19 (2004).
3. Sarker, A., Gonzalez, G.: Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics, 53, pp. 196-207 (2015).
4. Zhang, Y., Cui, S., Gao, H.: Adverse drug reaction detection on social media with deep linguistic features. Journal of biomedical informatics, 106, 103437 (2020).
5. Kwa, M., Welty, L. J., Xu, S.: Adverse events reported to the US Food and Drug Administration for cosmetics and personal care products. JAMA internal medicine, 177(8), pp. 1202-1204 (2017).
6. CFSAN Adverse Event Reporting System (CAERS) Data Web Posting. https://www.fda.gov/Food/ComplianceEnforcement/ucm494015.htm. Accessed January 18, 2023.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17

7. Wang, J., Yu, L. C., Zhang, X.: Explainable detection of adverse drug reaction with imbalanced data distribution. PLoS Computational Biology, 18(6), e1010144 (2022).
8. Alhuzali, H., Ananiadou, S.: Improving classification of adverse drug reactions through using sentiment analysis and transfer learning. In Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 339-347 (2019).
9. Gurulingappa, H., Mateen-Rajpu, A., Toldo, L.: Extraction of potential adverse drug events from medical case reports. Journal of biomedical semantics, 3(1), pp. 1-10 (2012).
10. Miranda, D. S.: Automated detection of adverse drug reactions in the biomedical literature using convolutional neural networks and biomedical word embeddings (2018).
11. Ding, P., Zhou, X., Zhang, X., Wang, J., Lei, Z.: An attentive neural sequence labeling model for adverse drug reactions mentions extraction. IEEE Access, 6, pp. 73305-73315 (2018).
12. Breden, A., Moore, L.: Detecting adverse drug reactions from twitter through domain-specific preprocessing and bert ensembling. arXiv preprint arXiv:2005.06634 (2020).
13. Sloane, R., Osanlou, O., Lewis, D., Bollegala, D., Maskell, S., Pirmohamed, M.: Social media and pharmacovigilance: a review of the opportunities and challenges. British journal of clinical pharmacology, 80(4), pp. 910-920 (2015).
14. Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., ..., Gonzalez, G.: Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. In Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing, pp. 1-8 (2014).
15. Taher, G.: E-commerce: advantages and limitations. International Journal of Academic Research in Accounting Finance and Management Sciences, 11(1), pp. 153-165 (2021).
16. Pattanayak, R. K., Kumar, V. S., Raman, K., Surya, M. M., Pooja, M. R.: E-commerce Application with Analytics for Pharmaceutical Industry. In Soft Computing for Security Applications, pp. 291-298. Springer, Singapore (2023).
17. Tay, E.: Evaluating Bayesian Hierarchical Models and Decision Criteria for the Detection of Adverse Events in Vaccine Clinical Trials (2022).
18. Freitas, A., Carvalho, A.: A tutorial on hierarchical classification with applications in bioinformatics. Research and trends in data mining technologies and applications, pp. 175-208 (2007).
19. Bisser, S.: Introduction to Azure Cognitive Services. In Microsoft Conversational AI Platform for Developers (pp. 67-140). Apress, Berkeley, CA (2021).
20. Satapathi, A., Mishra, A.: Build a Multilanguage Text Translator Using Azure Cognitive Services. In Developing Cloud-Native Solutions with Microsoft Azure and. NET (pp. 231-248). Apress, Berkeley, CA, (2023).
21. Wan, Y., Yang, B., Wong, D. F., Chao, L. S., Yao, L., Zhang, H., Chen, B.: Challenges of neural machine translation for short texts. Comp. Linguistics, 48(2), pp. 321-342 (2022).
22. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794 (2016).
23. Breiman, L.: Random forests. Machine learning, 45(1), pp. 5-32 (2001).
24. Shah, K., Patel, H., Sanghvi, D., Shah, M.: A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1), pp. 1-16 (2020).
25. Haumahu, J. P., Permana, S. D. H., Yaddarabullah, Y.: Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost). In IOP Conference Series: Materials Science and Engineering, Vol. 1098, No. 5, p. 052081. IOP Publishing (2021).

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_17