Received: 31 August 2021 | Revised: 17 February 2022 | Accepted: 21 February 2022

DOI: 10.1111/cts.13268

ARTICLE

Machine learning approach to identify adverse events in scientific biomedical literature

Sonja Wewering[1] | **Claudia Pietsch**[1] | **Marc Sumner**[2] | **Kornél Markó**[2] | **Anna-Theresa Lülf-Averhoff**[1] | **David Baehrens**[2]

[1]Scientific & Competitive Intelligence, Bayer AG, Wuppertal, Germany

[2]Averbis GmbH, Freiburg, Germany

Correspondence

Sonja Wewering, Bayer AG, Building 0459, 138, Wuppertal 42096, Germany.
Email: sonja.wewering@bayer.com

Funding information

This project was funded by the Bayer Life Science Collaboration Program (Project: detAEct)

Abstract

Monitoring the occurrence of adverse events in the scientific literature is a mandatory process in drug marketing surveillance. This is a very time-consuming and complex task to fulfill the compliance and, most importantly, to ensure patient safety. Therefore, a machine learning (ML) algorithm has been trained to support this manual intellectual review process, by automatically providing a classification of the literature articles into two types. An algorithm has been designed to automatically classify "relevant articles" which are reporting any kind of drug safety relevant information, and those which are not reporting an adverse drug reaction as "not relevant." The review process is consisted of many rules and aspects which needed to be taken into consideration. Therefore, for the training of the algorithm, thousands of documents from previous screenings have been used. After several iterations of adjustments and fine tuning, the ML approach is definitively a great achievement in pre-sorting the articles into "relevant" and "non-relevant" and supporting the intellectual review process.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Using machine learning (ML) to make decisions based on previous decisions is becoming more prominent in the digital world. However, to implement such a workflow in a very regulated field is a big challenge.

WHAT QUESTION DID THIS STUDY ADDRESS?

To what extend is it possible to replace human decisions needing intellectual input by ML?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

It shows that it is to a certain extent possible to detect drug safety-related information to the drugs in focus in written text. Furthermore, it combines the methodologies to show which technical solutions are best.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

Using ML in more and more processes will gain efficiency and will make drug discovery, drug development, and postmarketing surveillance more efficient and, most importantly, it will increase the patients' safety.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Clinical and Translational Science* published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics.

INTRODUCTION

A mandatory process in post-marketing surveillance for a pharmaceutical company is the monitoring of adverse events (AE) or adverse drug reactions and reporting them to the health authorities.[1]

There are various ways how AEs can be reported to a drug manufacturer, such as patients calling in to report, via clinical studies, by the reports of healthcare professionals or biomedical literature.[1,2] This article is only focused on the AE detection in biomedical literature. This is a mandatory process for every pharmaceutical company and consists of the AE identification from all published literature articles reported on marketed drugs. The active screening process in biomedical databases is necessary to fulfill the legal responsibilities and to ensure patient safety[1] by being aware of possible events and making this information available in the patient's information (e.g., package leaflets). To fulfill the compliance, it is important to review the newly published literature documents within a very few working days. This timely relevant process is accompanied by the ever-increasing number of literature articles published year by year.[3]

The most important source for scientific literature in the field of AE monitoring is *EMBASE* (e.g., it is also used as a source for medical literature monitoring of the European Medicines Agency).[1] The literature articles listed in *EMBASE* have more than quadrupled in the past 10 years.[4] Therefore, the pharmaceutical companies must deal with a huge number of documents which needed to be monitored and are ever increasing.

The intellectual review is a manual screening process, which is very time-consuming and, most importantly, needs to be done in a short period of time to guarantee the patient's safety. Therefore, it is a great benefit to support this manual literature review by using machine learning (ML) to enhance the correctness of the review and perform the review more quickly. The idea for this approach is to automatically classify the literature articles into two buckets, one that contains all "relevant" literature articles and the other one containing the "not relevant" ones. All screened literature articles from the past have been stored, and can therefore serve as a large training set for the algorithm. The approach was to test different configurations for the algorithm to identify the best method to classify the document collection. Because with the current state of the art it is not possible to detect a 100% correct classification with an ML method, literature review will still be a manual process supported by an algorithm. However, an ML approach could reduce the manual effort and a combination of intellectual screening and an automated process would increase the accuracy of the process. This would support the highest goal—of securing the patients safety. Therefore, this study addresses the question if an ML approach could identify AEs and classify literature documents as "relevant/irrelevant" for further processing.

METHODS

Data set

To train and test the learning algorithms examples were used from existing cases of literature review that were assessed intellectually in the past. In the previous intellectual assessment, abstract, title, and full text of each publication were manually reviewed by human experts to identify the relevant publications. The reviewers marked all publications as either relevant or not relevant. The process described here encompasses the regular download of newly published articles on *EMBASE.COM*. First, documents are identified which mention one or more company products by using a huge thesaurus containing all the marketed pharmaceutical products of Bayer. This is a first subset of documents and needs to be further evaluated. With text-mining and highlighting options, the documents are prepared for a detailed intellectual review. From the documents which mention a Bayer product, those are identified which mention an AE that might be related to a company product. These documents are named as "relevant" and stored in an in-house database for further processing. All other articles are "not relevant." They are only stored for quality control purposes, as "irrelevant documents" must be randomly selected to be re-evaluated.

Table 1 shows the number of articles used to train and test the algorithm. Only titles, abstracts, and keywords of the publications were used for the ML, whereas full texts were not available for the training or automatic classification. Table 2 gives an example of the content used from the literature articles.

TABLE 1 Numbers of documents used for the training and testing of the ML algorithm

Total numbers of example documents	**123,458**	**100%**
Relevant	21,186	17%
Not relevant	102,272	83%
Training documents	**111,110**	**90%**
Relevant	19,066	17%
Not relevant	92,044	83%
Test documents	**12,348**	**10%**
Relevant	2120	17%
Not relevant	10,228	83%

Abbreviation: ML, machine learning.

TABLE 2 The content fields in EMBASE available for training and example data for each field

EMBASE field	Content	Example
Title	The title of the publication	Aspirin induced asthma – A review
Abstract	The abstract of the publication	Aspirin (other nonsteroidal anti-inflammatory drugs) is contra-indicated for asthmatics...
Keywords	Indexing terms	Acetylsalicylic acid/w1 – adverse drug reaction – oral drug administration, ...
CAS No.	Registry number	Acetylsalicylic acid – 493-53-8 – 50-78-2 – 53,663-74-4 – 53,664-49-6 – 63,781-77-1, ...

Only 17% of the documents in the data set are marked as relevant, whereas the remaining 83% are not. To measure the effect of the imbalance on the ML results, a second, balanced training set was created with an equal number of relevant and irrelevant articles. Therefore, irrelevant articles were removed from the original training set at random, until there was roughly the same amount of relevant and irrelevant documents (i.e., 19,066 relevant and 20,455 irrelevant documents).

By using training examples from the existing literature review cases that were done by human reviewers previously, in this study, classification models have been trained to reproduce the manual decisions from the past automatically.

Each ML model was trained with a random subset of titles, abstracts, keywords, and CAS No. references (the training set). The remaining records (the test set) were categorized by the ML model automatically according to the status relevant/not relevant. Results from automated categorizations were compared with those of the reviews conducted by humans (Figure 1).

The performance of the ML model was evaluated based on the precision, recall, and F1-score of its predictions for each label "relevant" and "irrelevant." Results reported here are the performance of these metrics on the separate test set that was not used during training of the models.

Precision is the ratio of correct predictions to all predictions. For example, if a model classifies 100 records as "relevant," of which 80 are correct, the model has a precision of 0.8. Recall is the ratio of correct predictions to the total number of examples of the label. For example, if there are 100 records labeled as relevant by human reviewers and the ML model classifies 90 of these as "relevant," then the model has a recall of 0.9. The F1-score is the harmonic mean of precision and recall and is used to summarize the quality of the ML model in a single number.

Comparison of machine learning methods

To choose an ML algorithm for the setting, the speed and prediction quality of Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Bidirectional Encoder Representations for Transformers (BERT), DistilBERT, and AlBERT were tested.

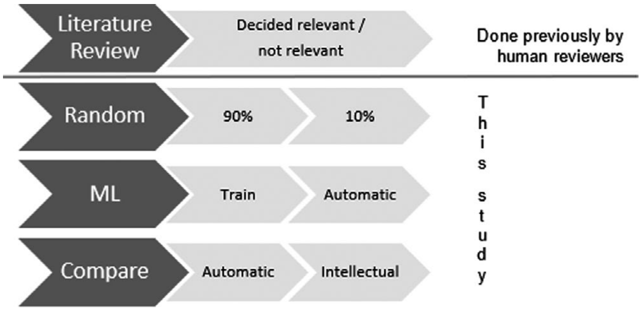

FIGURE 1 Approach to evaluate the results from automatic classification compared to the previous decisions of human reviewers. Remove 10% of the past decisions randomly from the examples and train the machine on the remaining 90% only. Then let the machine categorize the 10% automatically and compare with the intellectual decisions for the same. ML, machine learning

Machine learning algorithm used

The automatic classification has been realized with a combination of a well-proven ML methodology with advanced analysis functions. The fast SVM algorithm was used[5] and has been tuned for an optimized recall of those literature articles that are considered "relevant."

SVM is a supervised ML algorithm that is commonly used for the purpose of automatic classification.[6] SVMs are based on the idea of finding a linear separation (hyperplane) that divides a data set into two classes. To find the best separation, the algorithm chooses the hyperplane that results in the greatest distance between the hyperplane and the nearest data points (support vectors) from either training set of the two classes (maximum margin). By constructing more than one SVM, data can also be classified into three or more categories simultaneously (multilabel classification; Figure 2).

Tuning the SVM

The SVM was trained as a binary classifier which produces for each label ("relevant" or "irrelevant") a confidence value between zero and one. The label with the highest confidence value is taken to be the predicted label. The threshold for a prediction was set at 0.5, meaning the

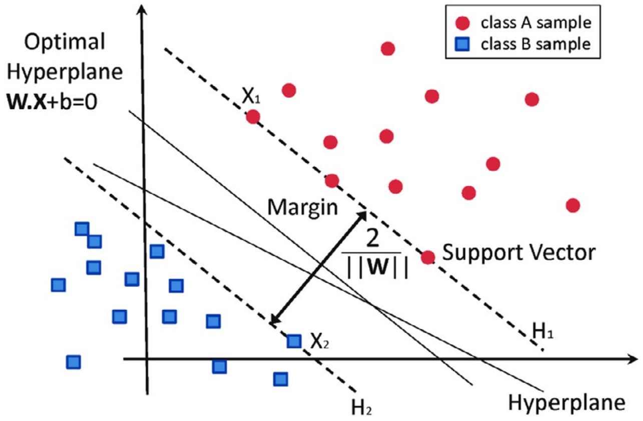

FIGURE 2 Classification of data by SVM.[7] SVM finds a linear separation (hyperplane) that divides the dataset into two classes. SVM, Support Vector Machine

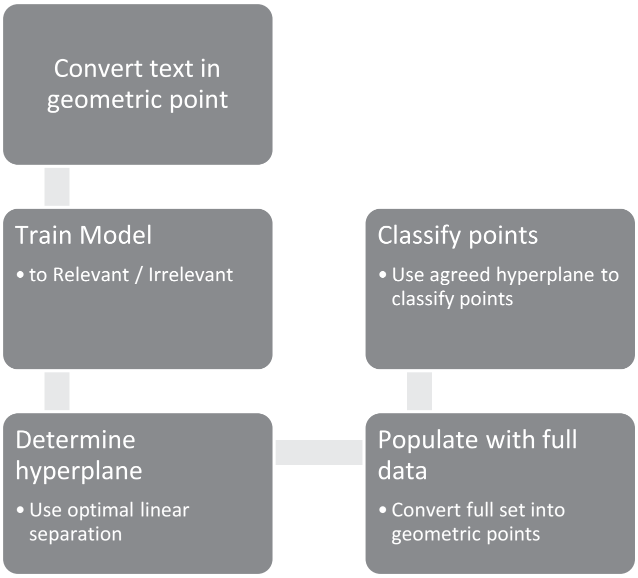
Convert text in geometric point
Train Model
• to Relevant / Irrelevant
Classify points
• Use agreed hyperplane to classify points
Determine hyperplane
• Use optimal linear separation
Populate with full data
• Convert full set into geometric points

FIGURE 3 SVM application to textual data. To apply SVMs to classify textual data, like the titles and abstracts of clinical studies, the texts need to be represented as points in a geometrical space (numerical vectors). SVM, Support Vector Machine

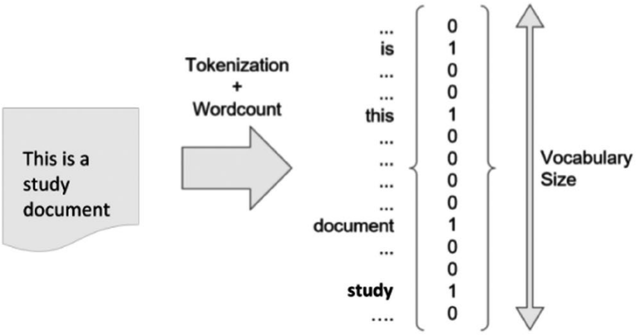

FIGURE 4 Convert text to geometric point with bag-of-words representation. Tokenization extracts words, stem forms of words are used, and counts of different stems

classifier always predicts a label (the sum of the confidence values is 1). It is possible to set this threshold higher, in which case the SVM would abstain from predicting a label if the highest confidence value is lower than the threshold.

Several models with differing thresholds based on a 10-fold cross-validation of the training data have been evaluated. Using a threshold of 0.5 produced the best trade-off between precision and recall. To achieve a higher recall of relevant documents, the balanced set of training data was used.

Text analysis

To apply SVMs to classify textual data like titles and abstracts of publications, the texts need to be represented as points in a geometrical space of numerical vectors, as shown in Figure 3.[8]

Therefore, the text of a study was decomposed into words (tokenization) and word counts with the same linguistic word stems (stemming) were added up in one entry of the vector representing the study text. The titles were processed separately from the abstract to allow the algorithm to give special weight to the title content. Tokenization was performed using the JTok library.[9] The JTok tokenizer is based on regular expressions with language-specific resources for abbreviations, punctuation, and clitics. The standard resources for English were used. Stemming was performed using the Snowball stemmer[10] for English. This is a rule-based approach to strip suffixes from tokens based on the Porter-Stemming algorithm.[11]

By this analysis, the counts of thousands of different word stems were derived for each study from the training set (bag-of-word representation). Thus, each document is a data point represented as a vector of stem counts (Figure 4) and the distance to a hyperplane can easily be computed. The distance of a data point from the separating hyperplane was used to compute a confidence value (0 to 1) for each label decision.

The software package Information Discovery[12] was used to perform the training of the SVM. The software includes the tokenizing and stemming algorithms necessary for the preprocessing of the texts. A single-label binary classifier with the default settings was used with the labels "relevant" and "irrelevant."

Improvement with synonyms?

More experiments were performed to answer the question if information from thesauri can be leveraged to improve the performance of the document classification.

For example, with original phrases in the studies like "… areas among pregnant women …" and "… Concerns over gestational effects …", does it help the algorithm to decide the relevancy if information is included about "pregnant" and "gestational" both refer to the same concept of pregnancy?

To find out, the SVM was trained with features expressing synonymity with concepts from two industry standard thesauri, MeSH and RxNorm, as well as some specialized thesauri curated by human experts in literature review for the purpose of pharmacovigilance.

RESULTS

Balanced data set

As stated in the Data set section, only 17% of the documents in the data set are marked relevant, whereas the remaining 83% are not. This leads to a bias in the SVM model to prefer classifying new documents as irrelevant. This is a known effect when working with imbalanced data sets and the SVM algorithm.[13]

To measure the effect of the imbalance of the original data set, two models were trained and classification performance was compared with the categorizations by the human reviewers using the separate test set described above. For the second model, the second set of training data were used, where relevant and irrelevant examples were balanced.

Table 3 shows the classification metrics of the two models on the test set. Although the overall performance is slightly higher for the imbalanced model (F1-score overall), the balanced model yields a significantly higher recall of relevant documents (recall relevant). However, the precision of the balanced model to recognize relevant documents is diminished (precision relevant; i.e., together with more relevant documents it also classifies more irrelevant documents as being relevant).

In addition to the assigned category, the confidence value of the SVMs was assessed to see if it is useful to improve on the reduced precision of the balanced training approach.

Figure S1 shows that it is possible to retrieve higher precision for the automatic classification of the relevant studies when requiring the confidence value to be above a given threshold (e.g., at least 90%; precision relevant). But with these thresholds at the same time the recall is reduced considerably (recall relevant). The harmonic mean of precision and recall is calculated by the F1-score and was found to stay almost the same for confidence thresholds between 60% and 80% (F1-score relevant). In conclusion, it is worthwhile to choose a threshold of 60% to achieve a high recall of 80% and maintain precision at 54%.

Comparison of machine learning methods

It was found that all methods had similar overall prediction quality on the data set (F1-scores between 0.85 and 0.88). DistilBERT and AlBERT had the highest portion of the documents predicted to be relevant that were also marked relevant by the human reviewers (precision of relevant class 0.72 for both). On the other hand, SVM identified the highest number of documents as relevant that were also marked relevant by the human reviewers (recall of relevant class 0.84). Computation time was the fastest with SVM (10 min for training, 1 min for prediction), whereas the other methods were considerably slower (50–300 min for training and 2–15 min for prediction).

Improvement with synonyms?

More experiments were performed to answer the question if information from thesauri can be leveraged to improve the performance of the document classification.

Table S1 shows the results from the experiments performed with synonym information from thesauri being available to the ML during training and prediction. It is observed that the synonyms do not add to the accuracy of the automatic classification on top of the word stems in the training data. This can be explained due to the many training examples covering already a variety of synonyms. The improvement in performance is minimal, at best, and does not justify the use of the thesauri information.

TABLE 3 Effect of relevant documents being the minority: Recall of relevant documents was improved by removing examples of documents marked irrelevant from the training data; however, this affects the precision of the model

Test set	F1-score overall	F1-score relevant	Precision relevant	Recall relevant	F1-score irrelevant	Precision irrelevant	Recall irrelevant
All data	0.8885	0.6317	0.7295	0.5571	0.9343	0.9125	0.9572
Balanced data	0.8279	0.6263	0.4993	**0.8401**	0.8882	0.9614	0.8254

DISCUSSION

For the described use-case, the goal was to use an algorithm to identify scientific literature articles in terms of the required Post-Launch Marketing Surveillance activities. The process requires to classify all documents in two "buckets"—relevant and irrelevant articles regarding the drug safety monitoring process. The primary goal was to detect all relevant literature articles (i.e., articles which report AEs to marketed products). Those documents need to be further analyzed in the pharmacovigilance process, because the most important aspect is to not miss any drug safety relevant documents in this process. With the method described here, it is possible to detect 84% of all relevant articles (see Table 3: Recall relevant = 0.84) if a balanced training set for the algorithm is used. Balancing the training set to contain equal numbers of relevant and irrelevant examples prevents the learning algorithm favoring the majority class of being irrelevant. It would allow to classify more irrelevant documents correctly than relevant ones incorrectly, resulting in a bias toward classifying as irrelevant wrongly. Our conclusion is that the use of thesaurus information is not worth the extra effort to include it to the ML model in our use case.

Balancing the training data achieves a high recall on unseen data; however, in terms of drug safety identifying all articles is a must and 100% correctness is needed. Therefore, until now, it is not possible to replace the intellectual screening efforts completely with an ML algorithm. But this automated classification will be very helpful in the screening process by doing a preselection of the documents. Intellectual screening processes can be focused on the irrelevant documents to identify those documents which are false negatives. This will support the screening process greatly in a timely manner.

The ML approach will be integrated into the daily work; therefore, the tool will be integrated into the information technology landscape by connecting it to our literature review system. As a first step, the ML tool will classify the documents automatically before the intellectual screening is started. Relevant documents with a high probability to be classified correctly (those with a high confidence value) will be further processed directly. The focus will be on documents with a low confidence value, for which the classification might be wrong. Here, the manual review will bring clarity. In total, focusing on documents with a low confidence score which need to be checked manually will save time in the review process. Besides, highly relevant documents are identified directly and can be analyzed further in terms of patient safety (e.g., reports to Heath Authorities, internal statistics, and possible adjustments in package leaflet). In addition, documents classified as irrelevant with a low confidence value need to be checked as well so that no relevant documents are missed. Furthermore, a combination of the automated process with the intellectual review will enhance the precision and will reduce mistakes in the process to reach nearly the 100% correctness.

In summary, the trained algorithm is a successful tool to support the intellectual screening process in terms of drug safety. It makes the process faster and less time-consuming and, most importantly, has an impact on patient safety, because the combination of ML and the manual screening enhances the precision not to miss important new safety information.

LIMITATIONS

The quality checks of the wrongly classified articles revealed one major issue when the ML is not classifying correctly. A major difference is that for the training of the algorithm only bibliographic data (title, abstract, and keywords; Table 2) can be used. The full texts are not available, due to license rights and availability on *EMBASE.COM*. For the intellectual review, the whole articles are, in some cases, ordered if more safety information is expected in the full text and if the decision on relevance cannot be made from title and abstract. This additional information is not available for the algorithm and those are the most cases and the major reason that the automatically classification is wrong. But on the other hand, it reveals how well the algorithm performs if enough textual information is available.

The quality of an SVM depends on the quality of the data on which it is trained. Although it can handle some noise in the data quite well, it will perform poorly if the human assessments in the training data are widely inconsistent. Additionally, an SVM needs labeled historical data to be able to train a model in the first place. Generally, for a binary classification problem, a few thousand records are sufficient.

ACKNOWLEDGEMENTS

The authors thank Guido Riewe for his intellectual input and critical review of the machine learning approach, and Dr. Peter Klügl and Louay Abdelgawad for their advice and support with the implementation and evaluation of the machine learning.

CONFLICT OF INTEREST

Sonja Wewering, Claudia Pietsch, and Anna-Theresa Lulf-Averhoff are employees of Bayer AG. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

S.W., D.B., and M.S. wrote the manuscript. S.W., C.P., K.M., A.L., and D.B. designed the research. D.B. and M.S. performed the research. S.W., C.P., D.B., and M.S. analyzed the data. D.B. and M.S. contributed new analytical tools.

ORCID

Sonja Wewering https://orcid.org/0000-0001-8625-6397

REFERENCES

1. European Medicines Agency. [Online] October 26, 2020. https://www.ema.europa.eu/en/human-regulatory/post-authorisation/pharmacovigilance/medical-literature-monitoring.
2. U.S: Food & Drug Administration. https://www.fda.gov/. [Online] 2021.
3. Scilit. *Market Size in Terms of Articles* [Online]. Scilit; 2021. [Cited: 7 16, 2021]. https://www.scilit.net/statistic-publishing-market-article.
4. Elsevier. *Embase content* [Online]. Embase/Elsevier; 2020. https://www.elsevier.com/solutions/embase-biomedical-research/embase-coverage-and-content#.
5. Joachims, T. Text categorization with support vector machines: learning with many relevant features. ECML-98 1398, 1998, pp. 137-142. 10.1007/BFb0026683
6. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: a library for large linear classification. *J Mach Learn Res.* 2008;61(9):1871-1874.
7. Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers – Scientific Figure on ResearchGate. s.l.: Available from: https://www.researchgate.net/figure/Classification-of-data-by-support-vector-machine-SVM_fig8_304611323. Accessed July 29, 2021.
8. Leopold, E., & Kindermann, J. Text categorization with support vector machines. How to represent texts in input space? *Mach Learn*. 2002, 46, 423-444. https://doi.org/10.1023/A:1012491419635
9. JTok. [Online] [Cited: 11 5, 2021]. https://github.com/DFKI-MLT/JTok.
10. Snowball [Online] [Cited: 11 5, 2021]. https://snowballstem.org/.
11. Porter MF. An algorithm for suffix stripping. *Dent Prog*. 1980;14:130-137.
12. Information Discovery. [Online] November 5, 2021. https://help.averbis.com/display/AKB/Information+Discovery.
13. Wu, G, & Change, E. *Class-Boundary Alignment for Imbalanced Dataset Learning*. Washington, DC: s.n. Computer Science; 2003. ICML 2003 Workshop on Learning from Imbalanced Data Sets II. pp. 49-56.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wewering S, Pietsch C, Sumner M, Markó K, Lülf-Averhoff A-T, Baehrens D. Machine learning approach to identify adverse events in scientific biomedical literature. *Clin Transl Sci*. 2022;00:1–7. doi:10.1111/cts.13268