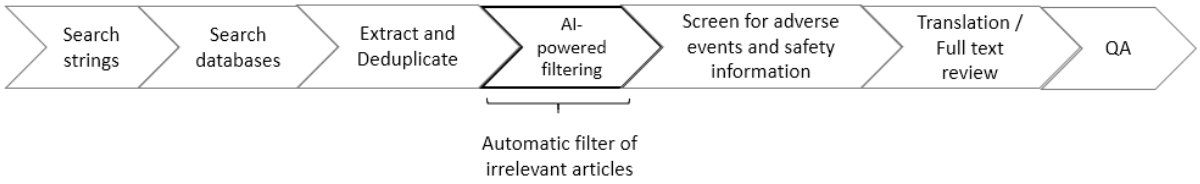
biologit

Model Fact Sheet

Adverse Event Detector | Rel. BETA2 | September 2021

1. Overview

The biologit adverse event detection system was developed to assist with high volume screening of scientific literature articles for safety events on medications. To achieve this goal the model is designed to be used as a first-pass ranking and filtering mechanism reflecting decisions made by pharmacovigilance specialists based on article abstracts following a typical literature monitoring process, as shown in the diagram below.



2. Intended Use

The system uses articles title and abstracts from the scientific literature as input and issues a prediction on whether the article contains one or more **suspect adverse event**. The model output enables an additional decision point to the screening process and the ability to rank or filter abstracts before reaching human specialists for further screening. The following uses were envisaged during the design of the model:

- Assist in the ranking and filtering of articles for the detection of individual case safety reports (ICSR) from the literature.
- Assist in the ranking and filtering of articles from the literature for the purposes of aggregate safety data reporting or signal detection.

2.1 Target Domains

The system is intended to be used on article abstracts from the scientific literature, specifically biomedical studies addressing a drug, compound or therapy to be used on human treatments.

2.2 Levels of Supervision

To support the validation and risk management requirements of pharmacovigilance processes, the system supports various degrees of human supervision, from human-in-the-loop processes with full-validation of results by specialists to higher levels of automation combined with human verification and auditing - ie. processes with human-supervising-the-loop and the ability to audit and override AI decisions. The desired level of supervision is determined by customer requirements. Please consult [1,3] for further details on implementing various levels of supervision of the adverse model using biologit MLM-AI.

2.3 Adaptiveness

Machine learning models implemented in the system can be described as *static*: the models are trained using training data curated in advance and labelled by pharmacovigilance specialists. The models do not dynamically learn from user input. Static AI systems are desirable from a validation standpoint as the models can closely follow current guidance and leverage existing methodologies for computer system validation [4].

2.4 Inputs and Outputs

Attribute	Expected Input
Input format	Title and abstract of a scientific article in plain text format (UTF8 encoding)
Domain	Biomedical text typical in academic articles from the scientific literature
Language	English

2.5 Operating Envelope

A number of rule-based safeguards are in place to prevent the model from making predictions on conditions not seen during training. The following scenarios are automatically flagged as suspect adverse event (so they are presented to users) irrespective of the model predictions:

- Abstracts in languages other than English.
- Articles with no abstract, containing only “errata” description, or containing only very short informative notes.
- Abstracts with a number of tokens greater than 1000 or less than 5.

Untested adjacent domains

Although the current model may present accurate results in domains related to human medicinal products, they are not yet represented in the training set. These domains may be included in future releases, however we currently recommend their use coupled with extensive validation. Examples include medical devices, cosmetics and neutraceuticals, and veterinary products.

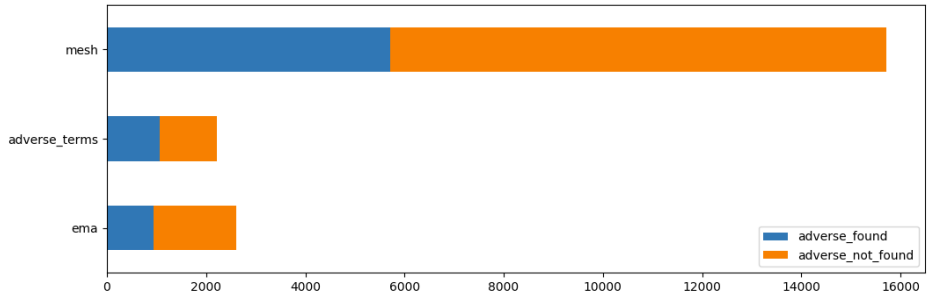
3. Training Data

The data set is built from a selection of publicly available articles (title and abstracts) from the scientific literature. To achieve broad coverage articles were selected based using pre-defined keyword searches that (1) reflected compounds belonging on one or more MeSH categories for therapeutic uses, (2) were part of the European Medicines Agency list of products under surveillance for medical literature monitoring, or (3) were retrieved using commonly used adverse terms or terms for special situations (pediatric, off-label, pregnancy-related terms, etc).

3.1 DATASET STATISTICS

Suspected Adverse Event	8,574
Not Suspected Adverse Event	14,239
Total Labels	22,813
Labels used in training (%)	20,531 (90%)

Distribution of training labels by article category:



3.2 Data Labelling Protocol

The labelling protocol is designed to reflect a typical literature monitoring workflow, where a first-pass decision is made upon screening abstract and title of a citation. In our protocol, an article is labelled as a suspected adverse event when (a) safety events are explicitly mentioned in the article abstract (applying to any product or treatment), or (b) contains implicit mentions of a safety event that may be fully described in the article full text (again independent of product or treatment).

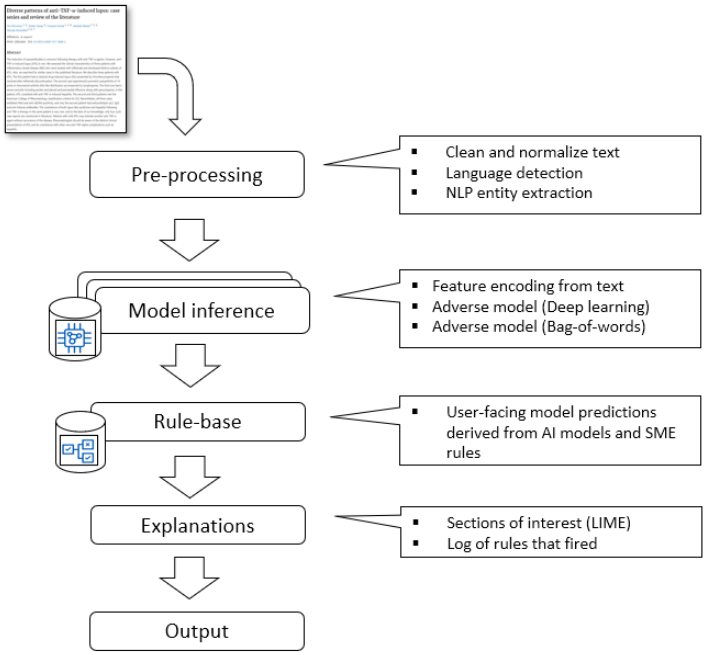
All labelled data undergoes sampled quality checks by a second annotator to ensure labels remain consistent. Inter-annotator agreement is monitored and additional revisions are triggered should it fall below an accepted threshold.

3.3 DATA SPLITS

The training set is built by randomly sampling 90% of all available labels (n=20,531), with remaining 10% split evenly into a validation set and a test set (n=1141). Sampling is stratified across the labelling categories discussed above. The training set is used to train the core machine learning models, while the validation set is used for hyper parameter tuning. The test set is not used in any training stages, and is set aside solely to report model performance results.

4. Model Description

The inference pipeline supporting the adverse model comprises a pre-processing stage where input text is cleaned and tokenized. This stage also performs language detection and a rule-based entity extraction of patient mentions which is used in later stages. The model inference stage encodes the normalized text into features and runs the prediction step of the machine learning models, producing raw model predictions. Next, a post-processing rules-based stage produces the final predictions and an explanation step computes additional metadata that can be used for helping users in understanding model predictions.



4.1 Machine Learning Models

Neural Model

The neural model employs a multi-layer neural network architecture organized as follows: an initial embedding layer converts tokens into vector representations using a combination of pre-trained word embeddings built with a biomedical text corpus [REF] and additional trainable embedding layers derived from part-of-speech tags and dependency parsing tags. The embeddings are combined and processed by a series of convolutional layers followed by a LSTM recurrent layer and an attention layer. Regularization is applied across the network architecture by using drop out during training and the use of batch normalization layers.

Bag-of-words Model

The neural model is supplemented by a bag-of-words model using 1-gram and 2-grams as features and trained with a random forest estimator.

Issuing Predictions

During the rule-based inference stage, the neural and bag-of-word model predictions are combined and subject to override rules authored in conjunction with pharmacovigilance subject matter experts.

4.2 Performance Metric

In drug safety, model mistakes have an asymmetric risk profile: articles falsely identified as a safety event (false positive) incurs incremental screening effort, while articles falsely identified as *not* a safety event (false negative) has a negative impact on what safety information is detected. Therefore *false negatives* are riskier and it is of paramount importance that this metric is minimized, even if at the expense of additional effort.

To ensure the rate of false negatives remains statistically within bounds, the adverse event model is parametrized for a desired target recall level. With desired recall fixed at a sufficiently high level, a metric that reflects the additional effort caused by false positives should be minimized. We use the *false positive rate*, defined as the ratio of false positives (FP) to the number of ground truth negative examples (N) given by:

$$\frac{FP}{N} = \frac{FP}{FP + TN}$$

Where FP is the number of false positives and TN is the number of true negatives. Thus, the performance target is the minimization of false positive rate at a desired target recall, set at 99%.

4.3 Experimental Results

Test set results for this release, tuned for a 99% desired recall are shown below. All metrics are with respect to suspect adverse found class.

Metric (adverse class)	Value
Recall	98.8%
False Positive Rate	45%
Precision	57%
f1 score	0.72

5. References

[1] Achieving Faster Literature Screening with AI - biologit Blog (November 2020).

[2] Reducing screening workload in medical literature monitoring with machine learning - DIA Regulatory Science Forum (September 2020).

[3] AI-based screening workflows - biologit MLM-AI product documentation.

[4] Huysentruyt, K., Kjoersvik, O., Dobracki, P. et al. *Validating Intelligent Automation Systems in Pharmacovigilance*, Drug Safety Issue 44 (2021) - https://doi.org/10.1007/s40264-020-01030-2