115632 – WEB-RADR

WEB- Recognising Adverse Drug Reactions

WP2B – Analytics

D2B.1 Technical report describing implementation and evaluation of AE recognition in social media

Lead contributor	Johan Ellenius (#12 – UMC)
	Johan.Ellenius@who-umc.org
Other contributors	Sara Hedfors Vidlin (#12 – UMC)
	Lucie Gattepaille (#12 – UMC)
	Tomas Bergvall (#12 – UMC)
	Carrie Pierce (#3 - Epidemico)

Due date	Month 18
Delivery date	Month 37 (September 2017)
Deliverable type	Report
Dissemination level	CO[1]

Description of Work	**Version**	**Date**
	V1.1	28-09-2017

[1] Please choose the appropriate reference and delete the rest:
PU = Public
PP = Restricted to other programme participants (including the Commission Services)
RE = Restricted to a group specified by the consortium (including the Commission Services)
CO = Confidential, only for members of the consortium (including the Commission Services)

Document History

Version	Date	Description
V1.0	21-09-2017	Draft for comments
V1.1	28-09-2017	Final Version

This report extends the M2B.3 report.

Summary

The aim of this report is to evaluate and present the results of the final evaluation of the benchmark analysis of refined methods for AE recognition in Twitter. It is an amendment to the previous deliverable M2B.3 from WP2b in the Web-RADR project.

The Web-RADR WP2b includes creating a new gold standard annotated Reference set. As this work stream is now finished, the new Reference dataset can be used to evaluate the methods developed in the work package. The original AE recognition pipeline described in deliverable M2B.3 report was applied to the Reference dataset. The *overall* classification performance of this pipeline has a recall of 0.09 and a precision equal to 0.50. The F-score was equal to 0.15.

Since the work presented in deliverable M2B.3 report identified the Medical Event (ME) Named Entity Recognition (NER) module as the major performance bottleneck in the pipeline, we have enhanced the NER module with two additional resources: an additional dictionary lookup based on a vernacular developed at Epidemico, mapping various expressions found in social media to a regulatory terminology (MedDRA Preferred Terms), and a logistic regression classifier trained to recognize the presence of MEs among 172 MedDRA PTs. To take into account these two additional ME NER resources, we have retrained an Adverse Event (AE) classifier for separating medicinal product / medical event pairs (MP/ME pairs) that are describing an adverse event from pairs where no such attribution is implied. This new AE classifier, together with the new NER module, form an enhanced AE recognition pipeline. Applied to the Reference dataset, the enhanced pipeline has a recall of 0.23, a precision of 0.35, and an F-score of 0.27.

We found that a relevance filter of posts, such as Epidemico's Indicator Score (IS), is beneficial for the pipeline, confirming an analogous conclusion obtained in the M2B.3 report. Removing the relevance filter from the pipeline increased the recall to 0.28 but at a higher cost on the precision, which dropped to 0.22, leading to a decreased F-score of 0.24.

To put our results into perspective we have compared our enhanced pipeline with Epidemico's already existing algorithm. Part of Epidemico's core activity is to provide social media monitoring for drug safety. The partly automatic / partly supervised work stream they have developed over the years represents the state-of-the-art when it comes to detection of adverse events in social media posts. Running the Reference dataset through the automated part of Epidemico's pipeline has a recall of 0.32 and a precision equal to 0.18. The F-score was equal to 0.23. The automated part of Epidemico's pipeline does not include an AE classifier at MP/ME pair level, instead it classifies posts as AE posts or non-AE posts and consider all possible MP/ME pairs within an AE post as a medicinal product / adverse event pair (MP/AE pair).

The somewhat subjective choice of the MedDRA PT for the medical event (e.g. fatigue vs somnolence for the text 'I am sleepy', or convulsion vs seizure for the text 'I had a fit') makes the task of encoding much trickier and affects the performance of the pipeline. We attempted to mitigated this effect by evaluating the match between the gold standard annotations and the results of the pipeline at the MedDRA Higher Level Term (HLT) level. Nonetheless, a detailed analysis of the outcomes of the enhanced pipeline revealed that about 28% of the false positives could actually be accepted as a valid interpretation of the text. These spurious false positives cannot be resolved by evaluating on the HLT level and evaluating the performance of the pipeline on even higher strata of the MedDRA hierarchy would not be appropriate because of a too high degree of generality.

In conclusion, we have developed an AE recognition pipeline that utilizes several dictionaries, vernaculars and NLP techniques, and that shows a slight improvement over the state-of-the-art automated pipeline developed by Epidemico. The automatic recognition of AEs in Twitter posts remains a challenging problem, in part owing to the difficulty of capturing the medical events mentioned in the posts, due to the richness of expressions employed and of course the complexity of language semantics in general. When giving as much importance to the precision as to the recall, we obtained a performance of 0.27 in F-score. Whether this performance is satisfactory for a use of the automated pipeline downstream is entirely dependent on the intended use of the MP/ME pairs produced by the pipeline.

Notations

ME	medical event
MP	medicinal product
AE	adverse event
NER	named entity recognition
LLT	lowest level term (MedDRA)
PT	preferred term (MedDRA)
HLT	high level term (MedDRA)
TP	true positive
FP	false positive
TN	true negative
FN	false negative
IS	indicator score

Introduction

Benchmark analysis of refined methods for AE recognition in social media

In short, the pipeline developed and described in deliverable M2B.3 report performs the following steps:

1. Extracts relevant posts
2. Identifies mentions of medical products (MP) in the posts and associates them to an entry in a drug dictionary (in our case WHODrug)
3. Identifies mentions of medical events (ME) in the posts and associate them to a MedDRA PT
4. Identifies MP/AE pairs, by classifying previously identified ME as either AE or non-AE, as well as identifying a related MP

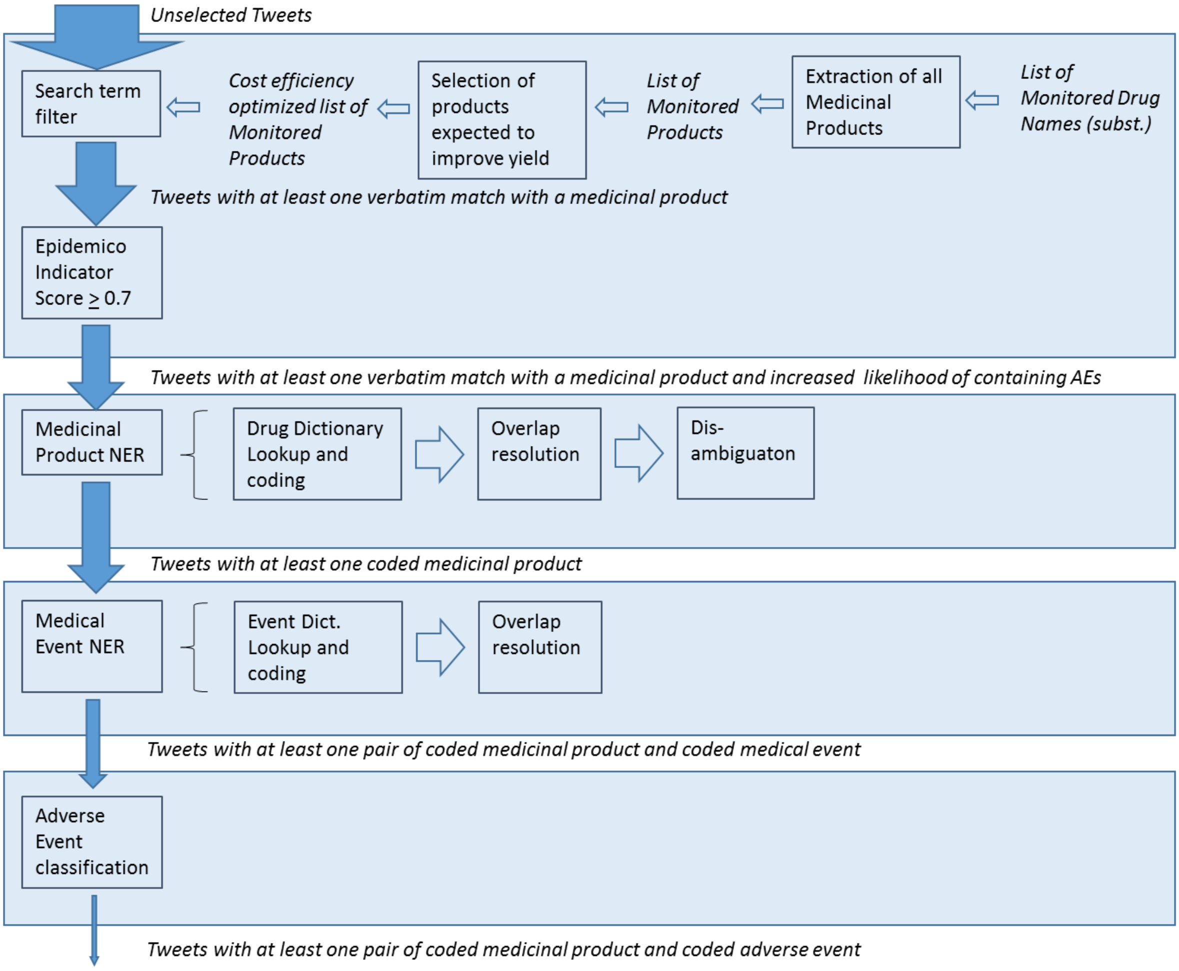
Unselected Tweets
Search term filter
Cost efficiency optimized list of Monitored Products
Selection of products expected to improve yield
List of Monitored Products
Extraction of all Medicinal Products
List of Monitored Drug Names (subst.)
Tweets with at least one verbatim match with a medicinal product
Epidemico Indicator Score ≥ 0.7
Tweets with at least one verbatim match with a medicinal product and increased likelihood of containing AEs
Medicinal Product NER
Drug Dictionary Lookup and coding
Overlap resolution
Dis-ambiguaton
Tweets with at least one coded medicinal product
Medical Event NER
Event Dict. Lookup and coding
Overlap resolution
Tweets with at least one pair of coded medicinal product and coded medical event
Adverse Event classification
Tweets with at least one pair of coded medicinal product and coded adverse event

Figure 1. *The processing pipeline with all computational components.*

Summary of the major findings presented in the M2B.3 report

1 – The main challenge in the recognition and encoding MP/AE pairs lies in the recognition and encoding of medical events. In particular, we have shown that by adding a vernacular derived from a free text field in VigiBase, where the reporter can describe a reaction, we can double the number of events detected without affecting the precision much, compared to using MedDRA LLTs as sole dictionary for dictionary lookup in Tweets. We hypothesized that adding more dictionaries and resources to capture medical events should improve the performance of this AE recognition pipeline.

2 – A relevance filter of posts, such as Epidemico's Indicator Score (IS) used with a decision threshold equal to 0.7, substantially improved precision but only marginally decreased recall. We thus concluded that such a filter should be included in the pipeline. In our previous technical report, this pipeline was named "track 4".

Research questions

The primary objective of this final evaluation of our developed methods is to prospectively assess the performance on a new gold standard annotated Reference dataset of Tweets.

- Is it possible to develop an AE recognition pipeline that achieve a performance that fulfils requirements expressed in terms of recall and precision?

As was mentioned above, Epidemico has already developed a dictionary based method for recognizing AEs in Twitter posts. An important research question is thus:

- How does the method developed in this study compare with the method developed by Epidemico?

Methods

Evaluation statistics

As in the deliverable M2B.3 report, we use a number of different statistics to evaluate performance: precision (P), recall (R), and F-score (F).

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F = \frac{2 * P * R}{P + R}$$

All MP/ME pairs found within a post referring to the same drug and MedDRA PT are aggregated, resulting in what we define as unique pairs. A MP/ME pair found and classified as a MP/AE pair is regarded as a TP if the gold standard annotation also has identified a MP/AE pair for that post where all the following are true:

1. The MP is mapped to the same substance in WHODrug
2. The AE is mapped to a PT in the same MedDRA HLT as the PT reported by the gold standard annotation

If either criterion is false, the pair is regarded a FP. If a MP/AE pair from the gold standard annotation is not matched by any pair found by the pipeline, it is considered as a FN. If several unique MP/AE pairs found by the pipeline match a single MP/AE pair in the gold standard, only one pair is counted as a TP, the remaining ones are counted as FPs.

Data preparation

The Reference dataset

Creating a new manually assessed Reference dataset has been one of the major objectives of the work in WP2b. The development of this dataset is presented in another deliverable of WP2b. The dataset contains 57,481 posts, where 1,058 of the posts contain at least one MP/AE mention.

Six substances; zolpidem, insulin glargine, levetiracetam, methylphenidate, sorafenib and terbinafine; were chosen for monitoring. A total of 873 product tradenames corresponding to these substances were then used to extract a dataset from Twitter, fetched from March 1st 2012 to March 1st 2015, resulting in a dataset containing 5,645,336 Tweets. From this set, 57,481 post were randomly extracted. 33,170 posts had an indicator score above 0.3 and those were manually curated by two independent teams, annotating all MP/AE pairs. 1,058 posts contained at least one MP/AE mention and a total of 1,423 pairs were found. Out of these, 1,398 pairs were considered unique, after aggregating all pairs in each post referring to the same drug and MedDRA PT.

The evaluation is based only on the above mentioned six substances. References to other substances are not included in our estimated measures of performance.

Additional resources of the enhanced pipeline

Compared to the original pipeline presented in the M2B.3 report, we have added two new resources to the medical event NER module, aiming at an increased recall, and re-trained an AE classifier on MP/ME pairs produced by the enhanced NER module, using the annotated dataset provided by Epidemico and described in the M2B.3 report. The development of the new resources was made without any utilization of the new Reference dataset. The three novelties of the enhanced pipeline are presented below.

Epidemico event vernacular

Epidemico has, during several years, developed a medical event vernacular, derived from annotations of medical events in Tweets. We have not been able to use the vernacular in previous evaluations, since they were performed on data that was partly used for creation of the vernacular. Since this bias does not exist for the new Reference dataset we wanted to investigate what this additional dictionary, developed and constantly enriched with new expressions thanks to continued manual curation of social media content, could add to the pipeline.

NLP NER

Using the data obtained from Epidemico and described in deliverable M2B.3, we have developed a new ME NER based on Natural Language Processing (NLP) techniques. In short, we have simultaneously trained 172 logistic regression classifiers, each designed for recognizing the presence of a unique MedDRA PT in the Tweets, using the Tweets represented as bags-of-grams (up to tri-grams). From the full set of 196,533 Tweets obtained from Epidemico, we have applied a simple text matching procedure to remove duplicates, then applied the de-duplication algorithm developed in the WP2b Record Linkage project (cf. Record Linkage technical report D2B.2), with threshold of 49, leaving a set of 163,878 supposedly original Tweets (among which 15,575 contain at least one MP/AE pair). A number of pre-processing steps has then been applied to the Tweets (English language filter, medicinal product masking, tokenization). The set of Tweets was then randomly divided into training (70%), validation (20%) and test (10%). Rare tokens that either did not appear or appeared less than 10 times in the training set were masked. Tweets were converted into bags-of-grams, a representation similar to the bag-of-words representation, but to which we added bi-grams and tri-grams. For each MedDRA PT present at least 20 times in the training set, we trained a logistic regression classifier on the Tweets as bags-of-grams. The overall performance of all classifiers combined has a performance of around 0.74 in recall and 0.72 in precision for both the validation and test sets.

Re-trained AE classifier

As the original AE classifier presented in track 4 of the M2B.3 report has been developed solely using MP/ME pairs produced by the original NER module based on MedDRA LLT and the VigiBase generated event dictionary, we had to re-train an AE classifier, so that MP/ME pairs from the additional two ME NER resources could be taken into account.

The new AE classifier is based on a logistic regression, using a large set of numeric features (document features such as number of user mentions, number of URLs, number of words, presence or absence of words belonging to given Word2Vec clusters; features related to the MP such as the number of tokens in the MP hit or the number of reports in VigiBase for the MP; features related to the ME such as the log-transformed frequency of the least common word in the ME hit or the number of tokens in the ME hit; features related to the MP/ME combination such as whether the ME hit occurs after the MP hit or the number of words between the MP hit and the ME hit). The detailed list of all the features can be found in the appendix of the M2B.3 report.

The new AE classifier was trained on the same training data that was used for training the NLP NER described above (including all preprocessing steps). Compared to the original AE classifier described in the M2B.3 report, we have not performed any feature selection procedure and used all 2525 numeric features found in the training set with the inclusion criterion for the Word2Vec cluster features of being observed at least 10 times in the training set. The logistic regression was regularized, at the level of 10^{-4} and optimized using the RMSprop algorithm with learning rate of 10^{-4} during 150 epochs. The classification boundary was chosen to maximize the F-score of the classifier on the validation set.

The two pipelines

The original pipeline consists of:

- A relevance filter based on Epidemico's indicator score. Only posts with indicator score greater than 0.7 are retained. Any gold standard annotated MP/AE pair appearing in a post with indicator score lower than 0.7 is counted as a false negative.
- The MP NER described in the M2B.3 report.
- A ME NER based on dictionary lookups using MedDRA LLTs as well as the VigiBase generated event dictionary described in the M2B.3 report.
- The original AE classifier described in the M2B.3 report.

The enhanced pipeline consists of:

- A relevance filter based on Epidemico's indicator score. Only posts with indicator score greater than 0.7 are retained. Any gold standard annotated MP/AE pair appearing in a post with indicator score lower than 0.7 is counted as a false negative.
- The MP NER described in the M2B.3 report.
- A ME NER based on dictionary lookups using Epidemico's vernacular, MedDRA LLTs and the VigiBase generated event dictionary, as well as the NLP NER described above.
- The re-trained AE classifier described above.

The automated part of Epidemico's algorithm for AE recognition

Epidemico provides a platform for monitoring health topics derived from online conversations across media sources. Monitoring drug safety concerns in social media chatter is part of their portfolio. To do so, they combine automated aggregation processing with manual curation to detect potential safety issues in near real-time. Their semi-supervised pipeline for AE recognition in Twitter and other social media sources represents therefore the state-of-the-art performance for the task and provides a good benchmark to compare the AE recognition pipeline developed in this work package to. However, since the supervised part of Epidemico's pipeline requires the involvement of expert annotators, we have only been able to compare our method to the automated part of their AE recognition pipeline. As our pipeline is also designed to run free of human supervision, the two approaches are comparable and it makes sense to restrict our comparison to the automated part of Epidemico's pipeline.

Epidemico's AE recognition pipeline is restricted to find posts with mentions of AEs and identify the MP(s) as well as ME(s) present in the posts. They do not make the coupling between medical events and medicinal products to identify which MP(s) and ME(s) are describing MP/AE pairs, as we do in this work. Thus, to allow for comparison between the AE recognition pipeline presented here and Epidemico's, we have assumed a systematic AE attribution of all possible combinations of MP(s) and ME(s) found in posts identified as AE posts by Epidemico's pipeline. We have then evaluated performance at the MedDRA HLT level, exactly as the pipelines presented in this report. Conversely, we have made a comparison between the automated part of Epidemico's AE recognition pipeline and our pipeline at the post level by classifying every post containing a positive MP/AE pair as an AE post, making our results comparable to how Epidemico's results should be interpreted in the first place.

Results and Discussion

Performance analysis of the two pipelines

The performance of the complete as well as various components making up the original and the enhanced pipelines are presented in the figure below.

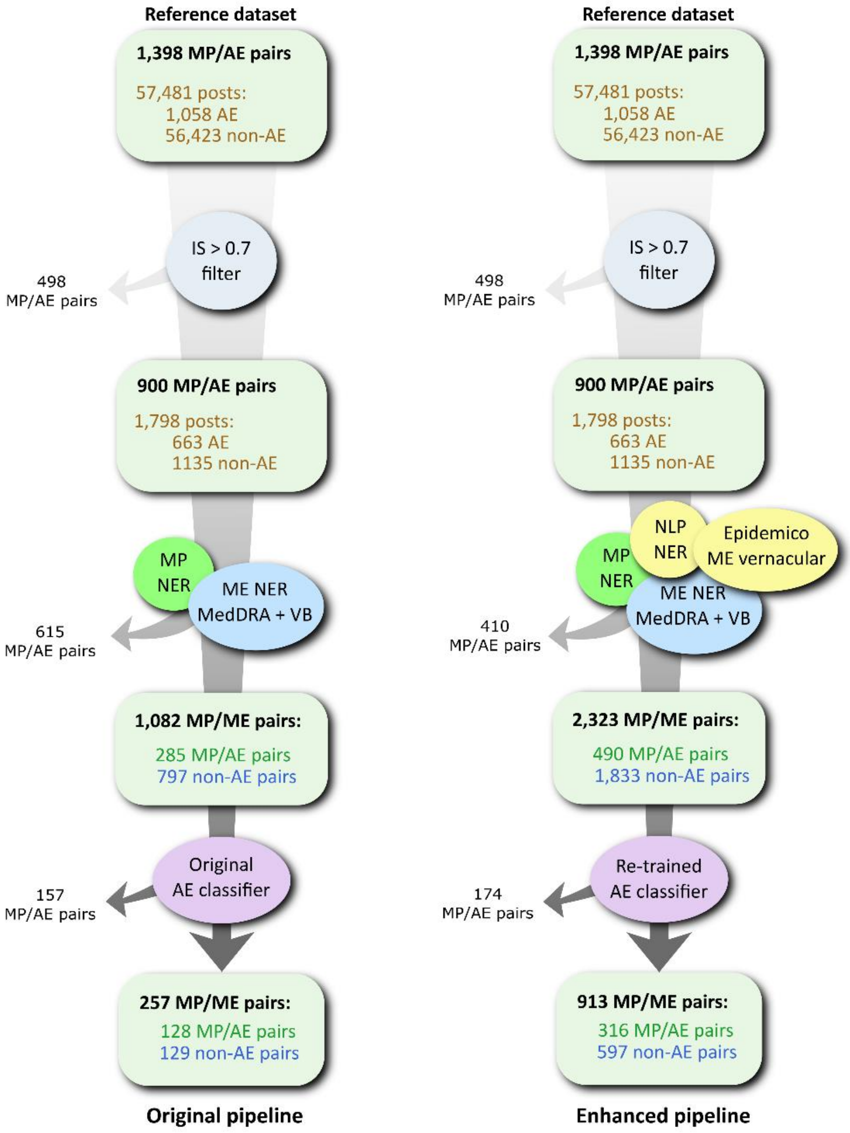

Figure 2. *Overview and performance of the original and enhanced pipelines.*

As we go along the pipeline, we lose MP/AE pairs, reducing the cumulative recall at every step. Table 1 presents the cumulative recall at various stages through the pipeline.

	Original pipeline			Enhanced pipeline		
Pipeline component	**Remaining MP/AE pairs**	**Component recall**	**Cumulative recall**	**Remaining MP/AE pairs**	**Component recall**	**Cumulative recall**
Start	1,398	1.00	1.00	1,398	1.00	1.00
IS filter	900	0.64	0.64	900	0.64	0.64
NER modules	285	0.32	0.20	490	0.54	0.35
AE classifier	128	0.45	0.09	316	0.64	0.23

Table 1. *Recall performance along the various components of the pipelines.*

On the previous dataset obtained from Epidemico and described in the M2B.3 report, the original pipeline gave the corresponding component recalls: 0.98 for the IS filter, 0.36 for the NER modules and 0.54 for the AE classifier. All component recalls have thus dropped when applied to the Reference set. The highest drop in recall is observed for the IS filter (from 0.98 on Epidemico's dataset to 0.64 on the Reference set) and is probably related to the fact that the parameters involved in computing the indicator scores are optimized using in-house data at Epidemico, that might include the Tweets used in the original evaluation done in the M2B.3 report. In contrast, the Reference dataset, except for filtering the Tweets selected for manual curation, has been developed independently of the IS. The loss of 34% of MP/AE pairs due to the IS filter might appear worrisome for the performance but, as we show in a section further down, is still beneficial because it removes more irrelevant posts than it filters relevant posts out.

The second component most affected by the change of dataset is the AE classifier. This also comes with no surprise, as it was trained on Epidemico's dataset (albeit evaluated on a disjoint part of the dataset of course) and should come with a certain degree of overfitting.

Finally, the NER modules of the original pipeline are the least affected by the change of dataset. Because they are based solely on dictionary lookups, they are not trained/fitted/tuned to any particular dataset, and therefore, their performance is expected to be dataset independent to a certain degree.

The enhanced pipeline clearly improves the recall over the original pipeline, as was intended by the addition of new ME NER resources and the re-training of an AE classifier. Nonetheless, the NER module (and in particular the ME NER module) remains the least performant component of our pipeline in terms of recall. Efforts should be made in developing more performant ways to improve the recognition and coding of MEs.

The table below presents the performances of the pipelines compared with the automated part of Epidemico's AE recognition pipeline.

Name	TP	FP	FN	Recall	Precision	F-score
Original pipeline	128	129	1,270	0.09	0.50	0.15
Enhanced pipeline	316	597	1,082	0.23	0.35	0.27
Epidemico's algorithm	443	2,005	955	0.32	0.18	0.23

Table 2. *Performances on the Reference set for recognizing MP/AE pairs. As a reference, the estimated performance of the original pipeline on the dataset made available to us by Epidemico and presented in M2B.3 had a recall, precision and F-score equal to 0.19, 0.44, and 0.27, respectively.*

We make the following observations:

- The enhanced pipeline has a much improved recall compared to the original pipeline (0.23 versus 0.09). The overall performance as measured by the F-score is also much improved (0.27 versus 0.15), even though the precision dropped somewhat (0.35 versus 0.50).
- Compared to the automated part of Epidemico's AE recognition pipeline, the enhanced pipeline has a higher F-score (0.27 versus 0.23). However, Epidemico's algorithm has a higher recall but a lower precision compared with the enhanced pipeline.
- More insights into how the enhanced pipeline compared with Epidemico's algorithm requires us to compare performances when tuned to equal levels of recall (or precision), performed later on in this report.
- The performance of the original pipeline on the previous dataset provided by Epidemico, compared with its performance on the Reference set presented in M2B.3, shows a significant drop in recall and F-score.

Origin of the MP/ME pairs in the enhanced pipeline

In the enhanced ME NER, MP/ME pairs can have 3 different origins: the original NER based on MedDRA LLT and VigiBase generated event dictionary, Epidemico's vernacular and the NLP NER. It is interesting to see how the MP/ME pairs produced by the ME NER are distributed among the three resources and how much overlap there is between them. The Venn diagram in the figure below illustrates the different origin of the MP/ME pairs produced by the ME NER module.

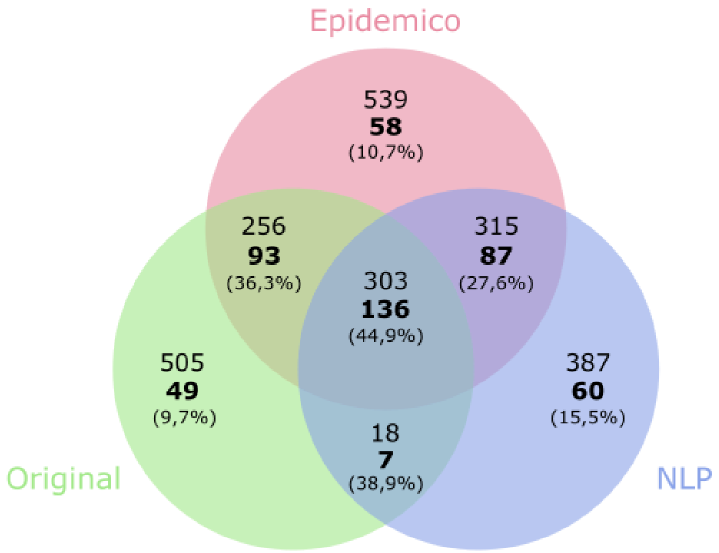

Figure 3. *Venn diagram showing the source of all MP/ME pairs produced by the NER modules in the enhanced pipeline. The numbers represent the total number of found MP/ME pairs including non-AEs (top), the number of found true MP/AE pairs according to the gold standard (bold), and the percentage of true MP/AE pairs (parenthesis).*

Each ME NER resource creates MP/AE pairs that are not found by the other NER components, indicating that all three resources are useful to get the recall as high as possible. However, should one of the three resources be abandoned, the original NER component based on MedDRA LLTs and the VigiBase generated event dictionary would represent the indicated choice, as it recognizes on its own the least amount of MP/AE pairs while producing a fair amount of MP/ME pairs (thus many false positives), although those numbers are rather similar to those concerning the Epidemico's dictionary. Remarkably, almost half of the MP/ME pairs found by all three resources are MP/AE pairs and for any MP/ME pair, being found by at least two of the resources triples the chance of being an MP/AE pair.

Classification performance and effect of relevance filter

The figure below presents classification performance for varying classification thresholds which facilitates comparison between the pipelines on either the same recall or precision. These figures also illustrate the effect of applying the IS filter.

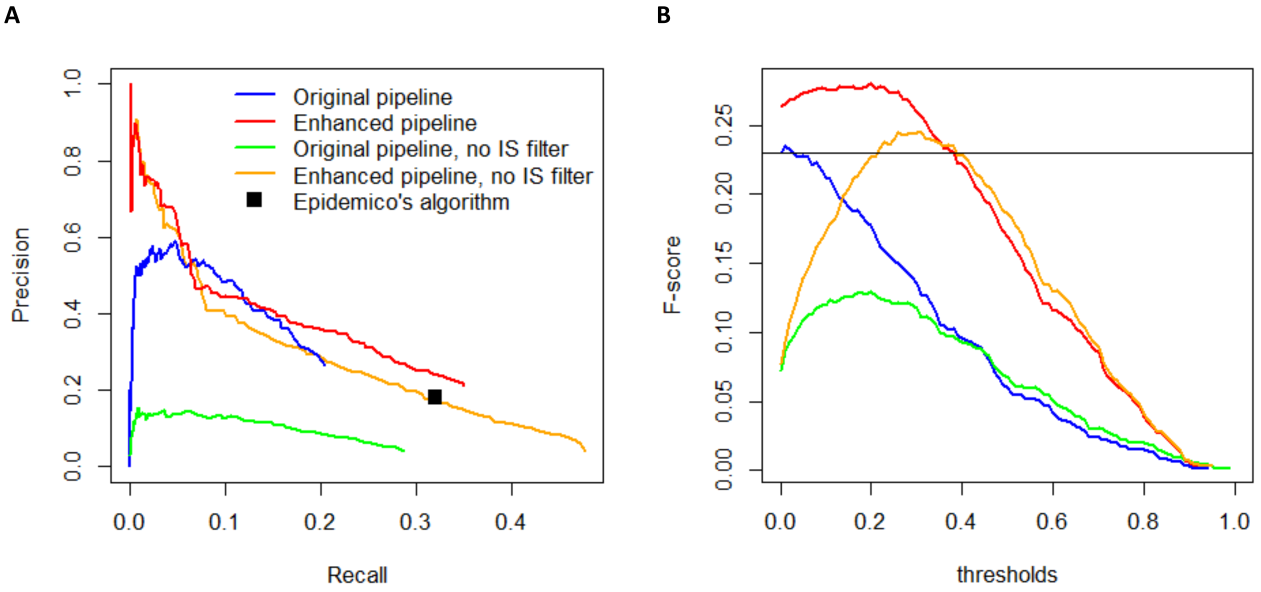

Figure 4. *Varying the decision threshold for the classifier with respect to precision and recall (A) and F-score (B).*

It can be noted that:

- The automated part of Epidemico’s AE recognition pipeline has a recall and precision equal to 0.32 and 0.18, respectively. If the enhanced pipeline is adapted to perform classifications with the same recall, its precision becomes equal to 0.24.
- Comparing the performance of the enhanced pipeline with Epidemico’s algorithm for the same precision is not possible since the precision of the enhanced pipeline is always greater than 0.18. However, a maximal recall of 0.35 is obtained when all produced MP/ME pairs are classified as MP/AE pairs, and the corresponding precision is 0.21 (F-score equals 0.26).
- For very low levels of recall (0.0 – 0.05), the enhanced pipeline offers much higher precision. This is however of only minor practical interest since recall should be significantly higher in order to be considered useful.
- In the recall interval 0.0-0.20, the utility of the IS filter when applied to the original pipeline becomes clear. It improves the precision with up to 0.09 absolute units for the enhanced pipeline and up to 0.45 absolute units for the original pipeline.

Comparison with Epidemico’s algorithm at the post level

As we mentioned earlier, the automated part of Epidemico’s AE recognition pipeline does not perform a formal attribution of medicinal products to the AEs they are involved with. To make a comparison to the pipeline developed in this work we have assumed a systematic attribution of all MPs to all MEs present in posts identified by Epidemico’s algorithm as AE posts. However, it is possible to compare the performances at the post level as well, ignoring the attribution of MPs to MEs in AE posts by classifying any post having at least one MP and one ME mentioned as an AE post. This AE-post classification does not consider the correctness of the encoding of the MPs and MEs in the posts. We obtain the results presented in the table below.

	Precision	Recall	F-score
Epidemico’s AE-post classification	0.36	0.62	0.46
Enhanced pipeline AE-post classification	0.72	0.40	0.50

Table 3. *Comparative performance of post-level classification of AEs.*

While Epidemico’s algorithm can identify more AE posts than the enhanced pipeline (recall of 0.62 against 0.40), the posts classified as AE-posts by the enhanced pipeline are twice as likely to be AE posts than posts classified as AE-posts by Epidemico’s algorithm. Depending on the intended use of the pipeline, both approaches might be preferable over the other, depending on whether we want to identify most of the posts containing AEs, at the cost of also extracting non-AE posts, or whether we want to guaranty that most identified posts are indeed AE-posts, at the cost of missing a number of AE posts. We discuss this issue in the following section.

On the practical use of the enhanced pipeline for AE recognition

In this sub stream, we developed and assessed the classification performance of a method to recognize AEs in Tweets. Given this, the subsequent step regards deciding whether the performance is good enough. This can be done by defining requirements of classification performance expressed in terms of recall and precision. Defining these values may be tricky, but should take into account the intended use of the method. We have identified two different use cases:

1. The method for AE recognition is used in an automated procedure to perform statistical signal detection without human intervention. This would require a high precision in order to avoid false positive MP/AEs pairs. However, favoring a high precision is usually associated with a lower recall. In our case, we have estimated that a required precision greater than 0.7 can be expected to lead to a very low recall, less than 0.05 (Figure 4). In other words, only 5% of all AEs would be captured which seems to be too low.
2. The method for AE recognition is used in a semi-automatic processing pipeline where human intervention is introduced to manually curate recognized AEs, similar to the approach used by Epidemico. This would allow a lower precision since false positives may be detected and discarded, not allowing them to influence the computations performed in the subsequent step of statistical signal detection. However, this will obviously increase costs due to the manual intervention needed. In our case, a recall equal to 0.23 would lead to a precision equal to 0.35 (Table 2). For an even higher recall, 0.35, when all captured medical events are classified as AEs, the precision would be 0.21. Whether this performance is acceptable can be debated. Curators would find that about 4 out of 5 recognized MP/AE pairs are in fact false positives, capturing only around one third of all AEs.

As an illustration, we use the Reference dataset including the six monitored substances of 57,481 Tweets, containing 1,398 MP/AE pairs, to roughly estimate the expected number of pairs that would have to be considered over one month when using the enhanced pipeline according to use case 2. We assume a recall and precision equal to 0.23 and 0.35, respectively. Using the length of the time period for when Tweets were retrieved and scaling it to one month, we estimate that around 157,000 Tweets will be captured initially, containing around 2,500 highlighted MP/ME pairs. This sample would contain around 875 actual MP/AE pairs and the rest would be false positives. We estimate that the time required to manually curate 100 of the automatically classified MP/AE pair might take 40 minutes. This translates to a workload equal to around 17 hours. Note that this approach would effectively discover false positives. But false negatives cannot be found, only 23% of all MP/AE pairs would be recognized.

In conclusion, our presented method for AE recognition does not have sufficient classification performance to be supporting automatic statistical signal detection of Twitter. It may have sufficient performance when connected with manual curation. Ultimately, whether it is worthwhile to monitor AEs in Twitter for signal detection depends on whether new signals can be detected this way, an issue investigated in another sub stream.

Characterization of outcomes in the pipeline

A characterization of various types of errors performed by the computational components is investigated by categorizing outcomes from classifier and the NER modules in the enhanced pipeline. The analysis shows, as presented in detail below, that the evaluation of the method is not straightforward. Even as we allow validation on HLT level there are synonyms or sufficiently similar PTs in different HLTs, for example PTs in the HLT *Emotional and mood disturbances NEC* and in *Mood disorders NEC*, generating errors in evaluation. We find that approximately 34% of the FPs should ideally have been TPs scrutinizing the PT found and the PT in gold standard. Similarly, 18% of the FNs from the NER modules should ideally have been TNs and 8% of the TNs should ideally have been FNs. This would theoretically have given us a performance of the enhanced pipeline of 0.57 in precision and 0.37 in recall. But in order to get the true performance we would need to manually curate the output.

In this current implementation, the enhanced pipeline does not resolve conflicts between the three ME NER resources, when they map the same part of a Tweet to different PTs. Instead, it creates several pairs for the same Tweet extract, which leads to FPs since the gold standard annotation would in most cases create a single MP/AE pair for the given part of the Tweet. The analysis of the outcomes showed that 66 of the in total 200 investigated pairs after the classification were affected by this. Developing a procedure to resolve conflicts between the pairs produced by the three different ME NER resources might reduce the amount of FPs created by the pipeline and improve the overall performance.

False positive AEs from classifier, (enhanced pipeline)

A random subset of 50 false positives MP/ME pairs incorrectly classified as AE were evaluated manually. These false positive errors made by the classifier can be divided into the following categories:

1. Genuine FP errors
    a. AE classification error (22%):
    The mentioned medical event is not an AE in this context, but wrongly classified as such.
    b. Pairing errors (12%):
    The pipeline finds a medical event that is indeed an AE but pairs it with the wrong medicinal product.
    c. Wrong event mapping (12%):
    The pipeline finds a correct AE mention in the text but misinterprets and/or miscodes it (such error is paired with a genuine false negative from NER modules, from category 1b).
    d. Not a medical event (10%):
    The pipeline extracts a mention that is not a medical event in this context.
    e. Several PTs are found for an AE (6%):
    The pipeline finds more than one PT for a medical event that is indeed an AE, all belonging to the same HLT as in the gold standard. But since gold standard only have one annotation, the first are evaluated as a genuine true positive and the other one as a genuine false positive.
2. Spurious FP errors
    a. Mismatches in coding of medicinal event (28%):
    The module finds the correct MP/AE mention but maps it slightly differently than the gold standard, but not incorrectly, so that it is falsely evaluated as a false positive (such error is paired with a spurious false negative from NER modules, from category 2a).
    b. Potentially missed AEs in gold standard (6%):
    The module finds a MP/AE mention not in the gold standard, that could be considered correct, making it is falsely evaluated as a false positive.
3. Unclear context
    a. Unclear what the Tweet actually means (4%).

Examples of the different categories of false positives are presented in the table below.

FP	Text		Outcome	Gold Standard
1a	Its 530 in the morning. Why am I as awake as I am, and in a good mood. Haven't taken my Concerta yet	MP: ME:	Concerta Insomnia (“why am i awake”) (HLT 10013510)	Not an AE
1a	Still speaking in grunts this morning. Methylin hasn't kicked in yet. #ADHD	MP: ME:	Methylin Drug ineffective (“methylin hasn't”) (HLT 10043409)	Not an AE
1b	@user Epilim made me suicidal... i got taken off it immediately! It was terrifying! And i'm on 3500 of keppra and still have fits! :(	MP: ME:	Keppra Suicidal ideation (“suicidal”) (HLT 10042459)	Not an AE
1b	@user Benedryl is not doing shit for me and I'm afraid of Ambien. I haven't had it in a long time. I need to not be near a blog	MP: ME:	Ambien Drug ineffective (“not doing shit”) (HLT 10043409)	Not an AE
1c	My teeth are starting to hurt. #nexavar	MP: ME:	Nexavar Pain (“to hurt”) (HLT 10033372)	Nexavar Toothache (HLT 10044049)
1c	@user zolpidem! Biggest rubbish ever I feel even worse! Thank you. Hope you & little man are okay xx	MP: ME:	Zolpidem Feeling abnormal (“feel worse”) (HLT 10068759)	Zolpidem Condition aggravated (HLT 10018072)

1d	Can't remember is anybody on keppra? May be on it soon. @user @user @user @user @user	MP: ME:	Keppra Memory impairment ("remember") (HLT 10001956, 10027177)	Not an AE
1d	Skipped my Lantus time about 2 hours ago cause I fell asleep, woke up so thirsty I can't even AND boom, im high.... Life of a diabetic...	MP: ME:	Lantus Somnolence ("fell asleep") (HLT 10013509, 10013980)	Not an AE
1e	I dunno why but on #stilnoct I get the feeling the visual hallucinations that appear seem more video-gamey	MP: ME1: ME2:	Stilnoct Hallucination, visual ("visual hallucinations") (HLT 10034374) Hallucination ("hallucinations") (HLT 10034374)	Stilnoct Hallucination, visual (HLT 10034374)
2a	Never taking concerta again. I almost died.	MP: ME:	Concerta Unevaluable event ("almost died") (HLT 10018072)	Concerta Adverse event (HLT 10043409)
2a	@user she's so amazin i know. if ur still loopy from stilnoct watch the video for 'beast' *__*	MP: ME:	Stilnoct Altered state of consciousness ("loopy") (HLT 10013509, 10027362)	Stilnoct Feeling abnormal (HLT 10068759)
2b	@user i used to get red bumps from Lantus so they switched me to Levemir. Might b worth looking into other options if it gets bad	MP: ME:	Lantus Therapy change ("switched me to") (HLT 10027700)	Not an AE
2b	Feel & look rough these dam levetiracetam making me soooooo tired all the time URL	MP: ME:	Levetiracetam Fatigue ("tired all the time") (HLT 10003550)	Not an AE
3a	So I started taking my concerta and now I'm super focused....on how much I'm shaking ??	MP: ME:	Concerta Tremor ("shaking") (HLT 10044566)	Not an AE
3a	Steve Lawrence: Question, as posted before I am on keppra, Oxtellar, and Topomax. Lately my appetite has not been... URL	MP: ME:	Keppra Decreased appetite ("appetite") (HLT 10003022, 10068759)	Not an AE

***Table 4**. Example Tweets containing various categories of False Positive classifications performed by the AE classifier, enhanced pipeline.*

Since we have three separate ME NERs, we also have some intermediately events referring to the same words or subset of words in the post, but mapped to different MedDRA PTs. 25 of the 50 investigated pairs had such overlapping pairs.

False negative AEs from classifier (enhanced pipeline)

A random subset of 50 false negative MP/ME pairs incorrectly classified as not an AE were evaluated manually. These false negative errors made by the classifier can be divided into the following categories:

1. Genuine FN errors
    a. Classification mistakes (96%):
    The classifier erroneously classifies the MP/AE mention as a non AE.
    b. No classification error, but resulting in a genuine FN (4%):
    The classifier correctly classifies the words found as a non-AE, but this non AE is mapped to the same PT or HLT as a true AE in the post that is not picked up by the NER modules. The result is that the found non AE and the not found AE, referring to the same HLT, are evaluated as a false negative. This should ideally have been resulting in one true negative and one false negative.

FN	**Text**		**Outcome**	**Gold Standard**
1a	@user I feel the same #bgnow 17.4 I'm soo high right now. Lantus increase by 2u. Bring on the hypos. #confused	MP: ME:	Lantus Blood glucose abnormal ("hypos") (HLT 10007217) classified not AE	Lantus Blood glucose increased (HLT 10007217)
1a	@user I know the feeling about the kepra I'm on 500 mg twice a day and I feel like a zombie. I've been on it for 5 years too.	MP: AE:	Kepra Feeling abnormal ("zombie") (HLT 10068759) classified not AE	Kepra Feeling abnormal (HLT 10068759)

1b	@user no change over always caused me headake a and feeling sick had more side affects wiv kepra	MP: Keppra AE: Drug ineffective ("no change") (HLT 10043409) classified not AE	Keppra Adverse drug reaction (HLT 10043409)

Table 5*. Example Tweets containing various categories of False Negative classifications performed by the AE classifier, enhanced pipeline.*

16 of the 50 investigated pairs are overlapping with other pairs where the events are referring to the same words (fully or partially) but mapped to different PTs.

False negative AEs from NER modules (enhanced pipeline)

A random subset of 50 false negative MP/ME pairs never picked up by the NER modules (including the Epidemico event vernacular and NLP NER) were evaluated manually. These false negative errors, that never made it to the classifier, can be divided into the following categories:

1. Genuine FN errors

   a. Dictionary lookups failed to pick up either the medicinal product or the event mention (60%):
   Dictionary lookup failed to pick it up due to no exact matchings in the dictionary or due to misspellings in the post of either the medicinal product or event. Out of these, the following events were most common:

   a. 'drug ineffective', 'drug effect decreased' or 'condition aggravated' (17% of the 1a errors)
   b. 'feeling abnormal' or 'abnormal behaviour' (10% of the 1a errors)
   c. 'insomnia', 'hypersomnia' or 'sleep disorder' (7% of the 1a errors)
   d. 'hallucination' or 'delusion' (7% of the 1a errors)
   e. 'adverse event' (7% of the 1a errors)

   b. Miscoded event (22%):
   The module finds the correct MP/AE mention in the text but misinterprets and/or miscodes it (such error is paired either with a genuine false positive, from category 1c, if the classifier classifies the miscoded pair as an AE, or with a genuine true negative, from category 1c, if the classifier classifies the miscoded pair as not AE).

2. Spurious FN errors

   a. Mismatches in coding of the medical event (18%):
   The module finds the correct MP/AE mention but maps it slightly differently than the gold standard, but not incorrectly (such error is paired with a spurious false positive if the classifier classifies it as a AE, from category 2a, or with a spurious true negative, from category 2b, if the classifier classifies it as a non-AE).

FN	Text	Outcome	Gold Standard
1aa	Ambien ditched me half way thru the nigh and I'm super hungry too	MP: ME:	Ambien Drug ineffective (HLT 10043409)
1ab	My keppra makes me feel like I've been hit by a train :(lets see how great work will go...	MP: ME:	Keppra Feeling abnormal (HLT 10068759)
1ac	Keppra is taking its toll on me. In the last 48 hours I have slept 28 hours and I still feel tired. #SeizuresSuck	MP: ME:	Keppra Hypersomnia (HLT 10013980,10028714)
1ad	I once got these sleeping pills from my doctor I think they were zolnox before they knocked me out they made words on posters move around	MP: ME:	Zolnox Hallucination, visual (HLT 10034374)
1ae	90-year-olds and Stillnox do not mix well together. What a freaking nightmare.	MP: ME:	Stilnox Adverse event (HLT 10043409)

1b	To make matters more depressing I've figured out that Kepra may be great from preventing epilepsy fits but in the morning it makes me sick	MP: ME:	Kepra Malaise (“makes me sick”) (HLT 10003550)	Kepra Nausea (HLT 10028817)
1b	Can I wake up? Maybe my ambien is still in full effect or am I actually living a nightmare	MP: ME:	Ambien Nightmare ("nightmare") (HLT 10033920)	Ambien Feeling abnormal (HLT 10068759)
2a	I feel like Keppra is doing worse than epillim. In a really shit mood	MP: ME:	Keppra Affect lability ("mood") (HLT 10001438)	Keppra Mood altered (HLT 10014556)
2a	@user @user Keppra the brand version because I have an allergic reaction to levetiracetam.	MP: ME:	Levetiracetam Allergic reaction (“allergic reaction”) (HLT 10027654)	Levetiracetam Drug hypersensitivity (HLT 10001737)

***Table 6**. Example Tweets containing various categories of False Negative classifications, error occurring in the NER modules, enhanced pipeline.*

True negative AEs from classifier (enhanced pipeline)

A random subset of 50 true negative MP/ME pairs from the classifier were evaluated manually. These true negatives can be divided into the following categories:

1. Genuine TNs
   a. Not medicinal event in this context (44%):
   The module correctly discards a MP/ME pair, that is not present in the gold standard, where the event could be a description of a medicinal event, but is not in this context
   b. Medicinal event but no AE (30%):
   The module correctly discards a MP/ME pair, that is not present in the gold standard, where the event is a genuine medicinal event but not a description of an AE or it is an AE, but not in relation to the medicinal product in the pair
   c. Miscoded adverse event (14%):
   The module finds a genuine MP/AE mention in the text but misinterprets and/or miscodes it, and then classifies it as a non-AE (such error is paired with a genuine false negative from NER modules, from category 1b, or with a spurious false negative from classifier, from category 2a).
   d. Medicinal product word (4%):
   The module correctly discards a MP/ME pair, that is not present in the Gold Standard, where the event also a medicinal product
2. Spurious TN
   a. Potentially missed AEs in the gold standard (4%):
   The module finds a MP/AE mention not in the sold standard, that could be considered correct, making it falsely evaluated as a true negative.
   b. Mismatches in coding of the medical event (4%):
   The module finds the correct MP/AE mention but maps it slightly differently than the gold standard (but not incorrectly) and classifies it as a non-AE (such error is paired with a spurious false negative from NER modules, from category 2a).

TN	Text		Outcome	Gold Standard
1a	Yay! Reducing my Terbinafine dose 2day to twice a week!Still on high steroid dose but hopefully me will get my #Mojo back now! :)	MP: ME:	Terbinafine Altered state of consciousness (“high”) (HLT 10013509, 10027362) classified not AE	Not an AE
1a	@user its like date raping yourself, take enough and not fall asleep is pure crazy no memory at all, #zolpidem	MP: ME:	Zolpidem Acute lymphocytic leukaemia (“all”) (HLT 10024290) classified not AE	Not an AE
1b	Thank god for stilnox... Muscular cramps at the base of my spine kept me up half the night. Meep :(	MP: ME:	Stilnox Abdominal pain (“cramps) (HLT 10017926) classified not AE	Not an AE

1b	@user Years ago, done fuck all. I've got stilnox now, quarter or half a tab works well when needed. No drowsiness in the morn either.	MP: ME:	Stilnox Somnolence (“drowsiness”) (HLT 10013509, 10013980) classified not AE	Not an AE
1c	To make matters more depressing I've figured out that Kepra may be great from preventing epilepsy fits but in the morning it makes me sick	MP: ME:	Kepra Fatigue (“makes me”) (HLT 10003550) classified not AE	Kepra Nausea (HLT 10028817)
1c	Aha I'm fucked. Pretty high on sublinox right now :)	MP: ME:	Sublinox Pyrexia (“high”) (HLT 10016286) classified not AE	Sublinox Euphoric mood (HLT 10014556)
1d	Warming the insulin up a little bit this evening. Last night it stung. 1st time using #Lantus	MP: ME:	Lantus Blood glucose abnormal (“insulin”) (HLT 10007217) classified not AE	Not an AE
2a	@user Struggling with a migraine but mood much improved since I stopped Keppra. :)	MP: ME:	Keppra Affective disorder (“mood”) (HLT 10027948) classified not AE	Not an AE
2b	My Ritalin has my blood pressure up so high, I can be barely upset and I'll turn bright red and my head starts pounding.	MP: ME:	Ritalin Blood pressure increased (“blood pressure”) (HLT 10047110) classified not AE	Ritalin Hypertension (HLT 10020774)

Table 7. Example Tweets containing various categories of True Negative classifications, enhanced pipeline.

18 of the 50 investigated pairs are overlapping with other pairs where the events are referring to the same words (fully or partially) but mapped to different PTs.

True positive AEs from classifier (enhanced pipeline)

A random subset of 50 true positive MP/AE pairs from the classifier were evaluated manually. These true positives can be divided into the following categories:

1. Genuine TPs
    a. MP/AE mention with correct PT (86%):
    The MP/AE mention is correctly classified as an AE, with the same PT as in the gold standard.
    b. MP/AE mention with correct HLT (14%):
    The MP/ME mention is correctly classified as an AE, with a PT in the same HLT as in the gold standard.

TP	**Text**		**Outcome**	**Gold Standard**
1a	@user I was addicted to #stilnox and it's a terrible, evil drug to get off. I went cold turkey and it's horrendous !	MP: ME:	Stilnox Drug withdrawal syndrome (“cold turkey”) (HLT 10068756, 10079102)	Stilnox Drug withdrawal syndrome (HLT 10068756, 10079102)
1a	@user while OS i discovered the perfect cocktail of 1 stilnox + 2 restivat, i slept brilliantly but was so groggy in the morning!	MP: ME:	Stilnox Somnolence (“groggy”) (HLT 10013509, 10013980)	Stilnox Somnolence (HLT 10013509, 10013980)
1b	@user Praying you feel better. Dr. should definitely have answers. I only had 2 bad seizures. On keppra now, still get scary auras!	MP: ME:	Keppra Drug ineffective (“still get”) (HLT 10043409)	Keppra Drug ineffective for unapproved indication (HLT 10043409, 10079146)
1b	Idk how many times I've made it clear that fuckin Concerta doesn't really work for me. Most of what it does is just make me not eat.	MP: ME:	Concerta Drug ineffective (“work for me”) (HLT 10043409)	Concerta Drug effect decreased (HLT 10043409)

Table 8. Example Tweets containing various categories of True Positive classifications, enhanced pipeline.

7 of the 50 investigated pairs are overlapping with other pairs where the events are referring to the same words (fully or partially) but mapped to different PTs.

Conclusion

In this amendment, we evaluate the performance of a developed pipeline for recognition and coding of suspected medicinal product / adverse event pairs. The recall was equal to 0.23 and precision 0.35, resulting in an F-score of 0.27, quite equivalent to the performance presented in the previous evaluation in deliverable M2B.3 report (recall, 0.19, precision 0.44, F-score 0.27). The performance appears poor, but is comparable with the state-of-the-art method in this field that we compared it with - Epidemico's existing method - using the same Reference dataset, elucidating the complexity and challenges approaching this problem.

Acknowledgements

We gratefully acknowledge Niklas Norén and Daniele Sartori for helpful comments and suggestions. We also thank the gold standard annotation team (Antoni Wisniewski, Juergen Dietrich, Magnus Lerch, Letitia Jiri, Britta Anne Grum) for the curation and quality assurance of the Reference dataset of Tweets used in this study.

Appendix

Definition of concepts

The following concepts are used throughout the report

- *Medicinal product*: Article 1.1 of EU Directive 2004/27/EC provides the definition of "medicinal product" (a) Any substance or combination of substances presented as having properties for treating or preventing disease in human beings; or (b) Any substance or combination of substances which may be used in or administered to human beings either with a view to restoring, correcting or modifying physiological functions by exerting a pharmacological, immunological or metabolic action, or to making a medical diagnosis.
- *Medical Event*: We use the definition of "untoward occurrence": "any abnormal sign, symptom, or laboratory test, or any syndromic combination of such abnormalities, any untoward or unplanned occurrence (for example an accident or unplanned pregnancy), or any unexpected deterioration in a concurrent illness" (Aronson & Ferner, 2005)
- *Adverse Event*: "any untoward medical occurrence associated with the use of a drug in humans, whether or not considered drug related" (Inman, 1981)
- *Adverse Drug Reaction*: "a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function"(World Health Organization & others, 1972).