115632 – WEB-RADR

WEB- Recognising Adverse Drug Reactions

WP2B – Analytics

D2B.2 Technical report describing implementation and evaluation of record linkage in social media

Lead contributor	Tomas Bergvall (#12 – UMC)
	Tomas.Bergvall@who-umc.org
Other contributors	Niklas Norén (#12 – UMC)
	Lucie Gattepaille (#12 – UMC)
	Sara Vidlin (#12 – UMC)
	Sara Gama (#1 – Novartis Pharma)
	Jiri Letitia (#18 – Amgen)

Due date	Month 32
Delivery date	Month 36
Deliverable type	Report
Dissemination level	PP[1]

[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)

Description of Work	Version	Date
	V0.3	28-06-2017

Document History

Version	Date	Description
V0.1	19 June 2017	First Draft
V0.2	26 June 2017	Comments from partners
V0.3	28 June 2017	Added publishable summary
V1.0	Aug 2017	Final Version

Publishable Summary

The purpose of this technical report is to describe the work that has been done by the WEB-RADR consortium to develop a method that can detect duplicated content in social media, specifically Twitter. A probabilistic record linkage method called vigiMatch that was previously developed by the Uppsala Monitoring Centre to detect duplicated individual case safety reports has been used and adjusted to work with the free text information of Twitter posts.

Using this method together with an active learning approach to progressively make the method better gave promising results. In a dataset of 13,820 Twitter posts the method classified 2,330 posts (17%) as being duplicates where we estimate that only 1-3% of these posts are not real duplicates. With this approach, we could estimate that we can detect about 30% of all the duplicates in this data. Some examples where the method incorrectly thinks that the posts are duplicates include e.g. retweets where the user quotes an old tweet but adds some other personal information like “that happened to me too” or when the text is very similar but list concepts that are very common like “I have a headache”.

Introduction

Effective use of social media to detect adverse drug reactions is challenging and requires reliable data. Among the different factors that can affect data quality, duplication represents a challenge that needs to be addressed for several reasons. First, social media monitoring has the potential to produce tremendous amounts of data. Such volumes can be a burden in storage, computation or review, thus it is advantageous to exclude redundant information to reduce the volume and at the same time increase the proportion of relevant data. Second, most analyses rely on the assumption that each unique event is described in a single record. Duplication violates this assumption and may lead to an over-estimation of the amount of evidence in support of a particular association. Since duplication affects events differently (*e.g.* the same piece of news can be relayed in many posts while a personal experience might be described only once), it may also bias the analysis.

Social media content is frequently duplicated. This is especially true in Twitter, where users are encouraged to spread content by retweeting to make their friends aware of a given piece of information. Some such retweets add a comment or additional information and may not be considered pure duplicates, whereas others simply relay the original information. Such duplicates should be simple to detect thanks to large textual overlap, the presence of the “RT” tag in the text and/or the presence of a re-tweet flag in the xml code representing the tweet. However, other types of duplicates can be more difficult to identify: a user can refer to a single event in multiple posts and multiple users can refer to the same event in different ways. Conversely, textually close descriptions are not necessarily duplicates but may represent distinct events of similar nature.

One of the overall aims of the WEB-RADR project is to develop technical tools for data mining of publicly available data shared on social media website like Twitter. The aim of this study on record linkage is to detect and later exclude all duplicated Twitter posts without excluding posts that are not duplicates. One important design choice is whether to perform record linkage before and/or after eliminating posts of less relevance for pharmacovigilance. This choice has some implications on the method to be used to detect duplicate posts. First, in terms of how to represent the data e.g. as a bag of words or using the semantic meaning of the text. Second, the volume of data is very different with a larger number of pairwise comparisons needed earlier in the process. The framework within which this study has been performed is one where duplicate detection based on textual similarity of the posts would be performed prior to any filtering, which could in turn be followed by duplicate detection based on the semantic content of the posts. The latter step has not been investigated in this study.

Definition of concepts

The following concepts are used throughout the report

- *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”(1).
- *Adverse Event*: “any untoward medical occurrence associated with the use of a drug in humans, whether or not considered drug related”(2)
- *Duplicate:* Two entities are considered as duplicates if they are referring to the same event.
- *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.

Methods

Data collection

To develop a data mining method data is needed to estimate the internal parameters of the method, usually called “training the method”. To evaluate how well the method performs, this data also need to be manually annotated to know what the ground truth is. This allow researchers to evaluate where the method makes correct and erroneous decisions. We collected data from the publicly available Twitter Application Programming Interface (API)(3) during 2 months, starting 27 September 2016. 23 drug substances were selected to represent a wide range of uses on social media. The selection included both drugs and vaccines from a wide range of therapeutic areas. The search terms used for data collection were based on the tradenames for these 23 substances identified in WHODrug, the world’s largest repository of global tradenames. To reduce the amount of irrelevant posts, tradenames with a higher chance of being used in other contexts than referring to a medicinal product were excluded using a pre-existing method(4). This reduced the list of search terms from 1,412 to 133.
A limitation of the Twitter API is that only nine days of historical data can be retrieved. Hence, an iterative procedure was created where the list of search terms was used repeatedly to find all related tweets in the given period. In total, 13,820 tweets where collected, with substantial variation between substances, e.g. 23% of the posts were for HPV vaccine. The appendix show more information about the number of tweets collected for each substance and tradename.

Data preparation

In Twitter, urls are automatically shortened and two identical urls are not guaranteed to have the same shortened url. Since this information could be important when comparing two tweets all urls where expanded to their original form. The retweets that could be identified either by the retweet flag in the xml code of the tweets or by the inclusion of an RT at the beginning of a tweet at the same time as having an identical text to another tweet in the reference set were removed.

Blocking

Even for the small dataset of 13,820 tweets in our training data, there are more than 95 million possible pairwise comparisons, which can lead to large computing times. To reduce the number of comparisons

made, we use a blocking technique where records are grouped together and pairwise comparisons are only made between records within the same block. We base the blocking on single words; every tweet containing a given word will be compared to all other tweets containing this specific word. We assume that if two tweets do not share any word in common, then they are unlikely to be duplicates and do not need to be compared. On the other hand, it means that tweets with large textual overlap will be compared multiple times. Despite the unnecessary repetition that this creates, blocking still reduces the number of comparisons made, from 95 million to 74 million.

vigiMatch

The basis for duplicate detection in WEB-RADR has been vigiMatch(5). vigiMatch is a method for duplicate detection using probabilistic record linkage. The method was originally developed for detecting duplicate reports in VigiBase, the World Health Organization's (WHO) global database of individual case safety reports. It is based on the hit-miss model for record linkage introduced by Copas & Hilton(6). vigiMatch provides a total match score for a pair of posts, using a weighted combination of contributions from the relevant fields. Matching information is always rewarded, and the size of the reward depends on how common the matching event is (greater rewards for matches on rare events) and mismatching information is penalized (greater penalties for fields that have been found to be less error-prone). Most of the parameters of vigiMatch are based on characteristics of the data set as a whole, such as the relative frequency of each covariate value, and only the probability that the observed value does not correspond to the true value (a *miss*) is estimated based on sets of confirmed duplicates. Another advantage of vigiMatch is that it accounts both for the accuracy and the richness of the matching information; thus, two posts with some mismatching information can receive a greater match score than two shorter posts that are identical. For WEB-RADR, we implemented vigiMatch to screen for duplicates prior to relevance filtering, based on comparing the textual content between posts, as bags of words.
Once a vigiMatch score has been computed for every possible pair of posts, a threshold is set to classify the pairs into suspected duplicates or non-duplicates. We computed a threshold using the estimated precision obtained from a set of annotated pairs of posts, setting the desired precision at 0.99. Further review of randomly selected pairs among predicted duplicates lead us to consider two additional thresholds. Finally, the last step of the algorithm is to link all suspected duplicates to their counterparts. Indeed, if posts A and B are duplicates, as are A and C, then B and C are duplicates as well, and the collection thus forms a cluster representing one single post. After this single link clustering step, we remove all linked suspected duplicates except for one. The one to keep is selected based on which one has the highest score it receives when compared to itself.

Active Learning

The collected tweets were not already annotated for duplicates and manual examination of randomly selected pairs would have yielded a very low proportion of true duplicate pairs. Therefore, we applied active learning. This is a semi-supervised procedure where the algorithm to be trained is used to select samples that are submitted for manual annotation and iteratively learn from the annotated samples. At each iteration of the active learning procedure, around 200 pairs were sampled and annotated from ten different ranges of vigiMatch scores.
As described above, the probability of a *miss* in vigiMatch needs to be estimated from a set of known duplicates. However, this set of duplicates identified through our active learning iterations are not a representative sample of all duplicates in the underlying data, since they were selected based on vigiMatch. This enrichment of the training data with true duplicates was accounted for during parameter estimation by creating a modified set of training data that included multiple copies of the annotated pairs in such proportions as to match the relative number of true duplicates in each range of vigiMatch scores.
As a starting point of the procedure, we used two manually identified duplicate pairs of tweets to

estimate the probability of a *miss* and then applied vigiMatch to the entire dataset with the estimated parameters. We normalize the obtained vigiMatch scores of pairs to range between 0 and 1 (lowest and highest observed scores respectively) and sort pairs into 10 equally spaced score bins. We then randomly select 20 pairs from each score bin (some bins have a total count lower than 20, we sample all pairs in such cases) and submit them for manual annotation. The annotations were done by one single assessor, following the annotation flow-chart below in Figure 1.

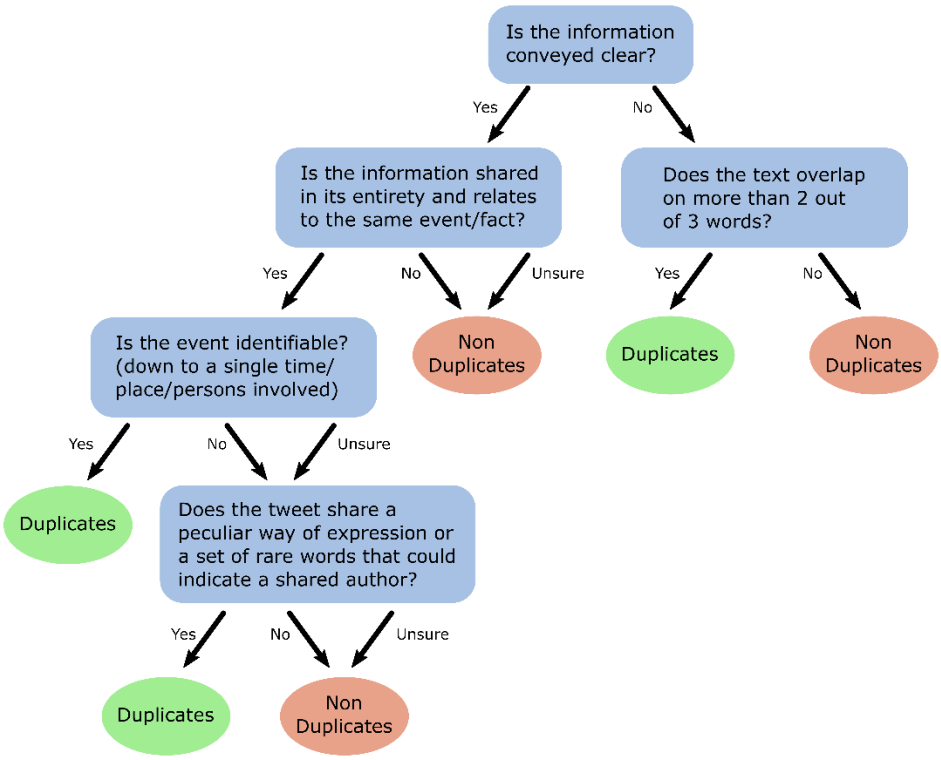

FIGURE 1: FLOW CHART OF THE ANNOTATION RULES

The pairs annotated as duplicates during the manual annotation are used in the subsequent active learning iteration to provide an updated estimate of the underlying parameters of vigiMatch. Because using every pair once would lead to a biased representation of the underlying distribution of scores within the class of duplicate pairs, we apply a re-sampling strategy based on the estimated proportion of duplicated pairs within each bin. Indeed, 1 out of 20 duplicate pairs in a bin containing a total 40,000 pairs suggests that the considered bin can have 2,000 duplicate pairs in total and hence should be represented more than the pairs in a bin where 20 out of 20 are annotated as duplicates but whose total number is 40. The re-sampling strategy aims at counterbalancing this representation bias. However, to avoid a single pair found in a dense score bin to drive the estimation of the parameters of vigiMatch, we did not sample directly according to the proportion of observed duplicates in each bin. Rather, we sampled the duplicate pairs with replacement according to the lower bound of the confidence interval for the estimated proportion, at the significance level of 0.01, multiplied by the total count of pairs in each bin. We illustrate this procedure using the two examples given above. We would sample the same duplicate pair 2.51e-4 * 40,000 ≈ 10 times in the case of the first example, while we would sample with replacement 0.767 * 40 ≈ 31 pairs from the 20 duplicate pairs in the case of the second example. If we would have used the estimated proportion directly instead of the lower bound of the confidence interval on the proportion, it would have resulted in 2,000 samples of the same duplicate pairs in the first example against 40 pairs in the second example, leading to a single pair driving the subsequent estimation of vigiMatch parameters. Once the input duplicate pairs for the next iteration are sampled, we repeat the 3-steps procedure described above. In total, we made 6 active learning iterations.

Evaluation

To evaluate the algorithm, we use the set of pairs that we manually annotated during the active learning procedure. This represents 974 pairs that were ascertained using random sampling in vigiMatch score bins throughout active learning iterations. As the vast majority of the pairs we have in our dataset are not annotated for duplication, we cannot provide the exact values of the recall and the precision of the algorithm for the entire dataset. However, by sampling with replacement from the annotated pairs according to the total number of pairs within each score bin at the last iteration, we can use the annotated data to estimate the performance the algorithm would have on the entire dataset.
We also directly evaluate randomly sampled pairs from predicted duplicates in order to get additional estimates of the precision. We sample 400 random pairs for 3 different vigiMatch score thresholds: 19, 35, and 50. We annotate them according to the annotation rules and record the number of observed false positive pairs.
To examine the added value of probabilistic record linkage over existing methods, vigiMatch was also run on already de-duplicated data provided by Epidemico. The data set consisted of 155,000 posts that had been manually annotated with possible adverse events and had been already screened for likely duplicates based on retweet flags and a Bloom filter(7). Any suspected duplicates should already have been eliminated so the data set contained no annotated duplicates, and was expected to contain a low proportion of duplicates, if any. We estimated the precision of our vigiMatch implementation through manual review of posts flagged as predicted duplicates. We reviewed predicted duplicate pairs where at least one post was annotated as a proto-AE, at 3 different thresholds of vigiMatch scores: 19 (400 randomly sampled pairs), 35 and 50 (all predicted duplicate pairs reviewed).

Results

Active Learning Iterations

We stopped the active learning procedure after 6 iterations. We obtained the following results over iterations (Table 1):

TABLE 1: ACTIVE LEARNING ITERATIONS

Iteration	# annotated duplicate pairs at the start	# annotated duplicate pairs at the end	Total number annotated pairs	*a**	*c**
1	2	100	188	0.088	0.169
2	87	162	345	0.084	0.161
3	148	227	508	0.075	0.144
4	209	288	665	0.072	0.138
5	267	348	818	0.079	0.151
6	326	411	974	0.074	0.142

*Note: *a* and *c* are vigiMatch parameters, as described in [Norén et al 2007].

After Iteration 3 no significant changes were found which was the reason for stopping at Iteration 6.

Evaluation

As shown in Figure 2, most pairs of posts are very different only a very small percentage of all pairs are predicted to be duplicates.

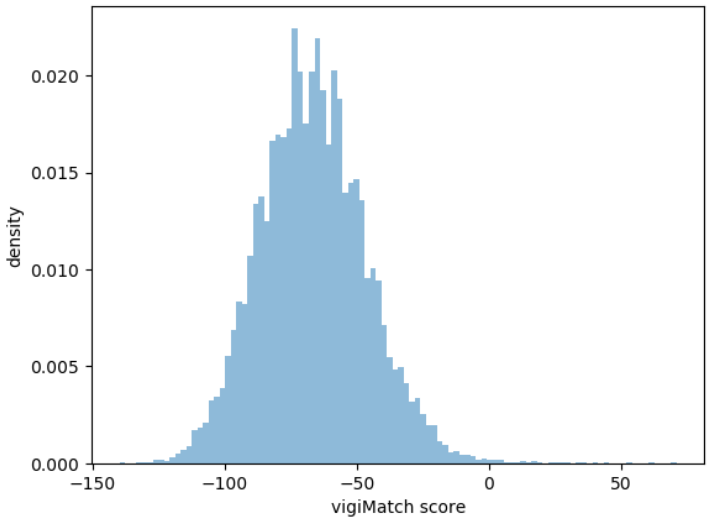

FIGURE 2: DISTRIBUTION OF VIGIMATCH SCORES FOR 20000 RANDOMLY SAMPLED PAIRS OF POSTS

The estimated performance of the algorithm on the entire dataset is provided in Figure 3 as a function of the classification threshold of vigiMatch scores. The vigiMatch scores in this dataset ranges between -169 and 223. The two metrics we are focusing the evaluation on are precision and recall.

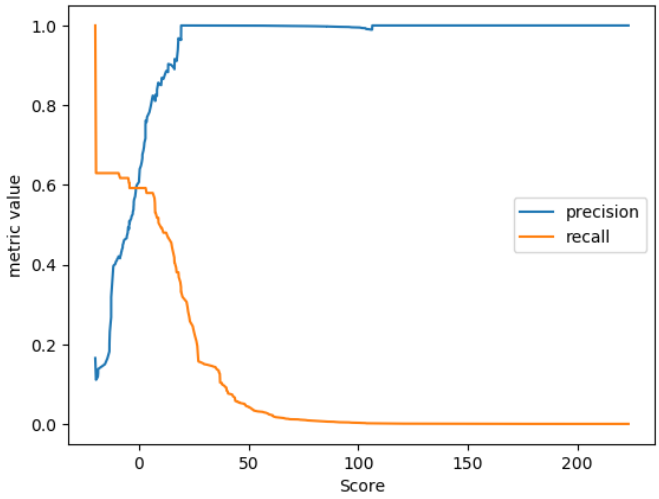

FIGURE 3: ESTIMATED PRECISION AND RECALL AS A FUNCTION OF THE SCORE THRESHOLD

The precision estimates in Figure 3 are computed using sampling with replacement according to the different size of score bins and the already annotated pairs obtained in the active learning procedure. Setting the desired precision to 0.99 or above, with best recall possible, we obtain a score threshold of 19. We compute then an additional precision estimate by randomly sampling 400 pairs with vigiMatch score above 19. Additionally, we randomly sampled 400 pairs with score above 35 and 400 pairs with score above 50. The results are given in Table 2Table 2.

TABLE 2 ESTIMATED PRECISION AT DIFFERENT SCORE THRESHOLDS ON THE TRAINING DATASET

Score threshold	Number of predicted duplicate pairs reviewed	Estimated precision
19	400 (out of 45,239)	95%
35	400 (out of 23,241)	97%
50	400 (out of 7,578)	98%

Performing the single link clustering step, we found that at a score threshold of 50, 2,330 posts are identified as duplicates (17%), whereas this number jumps to 5,584 (40%) when lowering the threshold to 19. Depending on the task at hand, the focus should be made on either the precision or the recall. In our case, as the de-duplicated data might be used for further analysis, we would probably want

to stay conservative and set a high threshold, to only remove posts for which we are confident of their duplicate status. This is especially important in problems where relevant information is rare and can share similarities, which is the case for the detection of AE posts. We would not want to remove posts reading "I took my medicine and now I have a headache" as they can be genuine accounts of separate events, despite being expressed in the exact same way.
To test the generalization of these results a dataset provided by Epidemico was used containing older Twitter data. In Table 3 the results after manual annotation are presented. Using the manually set threshold of 19 clearly did not give as good results on this dataset as expected. However, the original estimate of the threshold, 50, did provide better results.

TABLE 3. ANNOTATION RESULTS BASED ON THE EVALUATION DATASET

Score threshold	Number of predicted duplicates reviewed	Estimated precision
19	400 (out of 2,988)	14%
35	290 (out of 290)	91%
50	260 (out of 260)	97%

After the single link cluster analysis, we found 14,121 posts out of the 156,123 tweets (9%) in the dataset classified as duplicates. Looking at only the tweets containing adverse events this number drops to 1.5%.

Analysis of errors

Examples of the manual annotation presented in Table 2 and Table 3 are presented below.

UMC Twitter dataset with threshold at 19		
TP	Novo Nordisk receives Complete Response Letter in the US for faster-acting insulin aspart http://www.businesspress24.com/pressrelease1462697.html?utm_source=dlvr.it&utm_medium=twitter	Novo Nordisk receives Complete Response Letter in the US for fasteracting insulin aspart http://www.bioportfolio.com/news/article/2865966/Novo-Nordisk-receives-Complete-Response-Letter-in-the-US-for-faster-acting.html
	Pediatricians warn about toxic effects of Gardasil vaccine http://wp.me/p1Ue8F-25I	Pediatricians warn about toxic effects of Gardasil vaccine http://wp.me/p1Ue8F-25I
	Xgeva May Boost Bone Erosion Repair in RA http://www.medpagetoday.com/Rheumatology/Arthritis/61128 #arthritis	Xgeva May Boost Bone Erosion Repair in RA (CME/CE) #Rheumatology http://www.medpagetoday.com/Rheumatology/Arthritis/61128
	Gardasil propelled Merck to a strong third quarter — but don't expect that to last: New CDC guidelines should... http://www.marketwatch.com/story/gardasil-propelled-merck-to-a-strong-third-quarter-but-dont-expect-that-to-last-2016-10-25?siteid=rss&rss=1&utm_source=twitterfeed&utm_medium=twitter	Gardasil propelled Merck to a strong third quarter — but don't expect that to last: New CDC guidelines… http://www.marketwatch.com/story/gardasil-propelled-merck-to-a-strong-third-quarter-but-dont-expect-that-to-last-2016-10-25?siteid=rss&rss=1 MARKETWATCH
	Cetirizine, your fever's gripped me again Never kisses, all you ever send are full-stops	@jacksonbollocks cetirizine your fever's gripped me again never kisses all you ever send are full-stops
FP	zometa dosing &amp[semicolon] zometa infusion &amp[semicolon] zometa package insert &amp[semicolon] zometa cost &amp[semicolon] what is zometa &amp[semicolon] zometa side effects https://supremeupload.net/shrt/404.php	zoledronic acid cost &amp[semicolon] zoledronic acid package insert &amp[semicolon] zoledronic acid dosing &amp[semicolon] what is zoledronic acid https://supremeupload.net/shrt/404.php
	what is risperdal &amp[semicolon] risperdal weight gain &amp[semicolon] risperdal lawyers &amp[semicolon] risperdal consta &amp[semicolon] risperdal injection https://supremeupload.net/shrt/404.php	risperdal consta &amp[semicolon] risperdal medication &amp[semicolon] what is risperdal for &amp[semicolon] risperdal injection &amp[semicolon] what is risperdal https://supremeupload.net/shrt/404.php
	Off to hospital shortly for CT scan and monthly Zometa infusion. — feeling anxious	Off to hospital for monthly Zometa infusion.
	This fox needs Cipro!	This chipmunk needs Cipro!
	I added a video to a @YouTube playlist https://www.youtube.com/watch?v=gc7-0Vlw1pY&feature=youtu.be&a Guard Yourself A Gardasil Documentary Short	Guard yourself : A Gardasil documentary https://www.youtube.com/watch?v=piAXEafuM1c&sns=tw via @youtube

UMC Twitter dataset with threshold at 50		
TP	Ivacaftor and symptoms of extra-oesophageal reflux in patients with cystic… http://www.sciencedirect.com/science/article/pii/S1569199316305616 #eprompt #respire	Ivacaftor and symptoms of extra-oesophageal reflux in patients with cystic… http://www.sciencedirect.com/science/article/pii/S1569199316305616 #eprompt #respire
	Merck beats estimates following better-than-expected sales of HPV vaccine https://www.bloomberg.com/news/articles/2016-10-25/merck-beats-estimates-as-sales-of-gardasil-vaccine-surge?utm_content=business&cmpid=socialflow-twitter-business&utm_campaign=socialflow-organic&utm_source=twitter&utm_medium=social https://twitter.com/business/status/790874936684470272/photo/1	business: Merck beats estimates following better-than-expected sales of HPV vaccine https://www.bloomberg.com/news/articles/2016-10-25/merck-beats-estimates-as-sales-of-gardasil-vaccine-surge?utm_content=business&cmpid=socialflow-twitter-business&utm_campaign=socialflow-organic&utm_source=twitter&utm_medium=social https://twitter.com/business/status/790874936684470272/photo/1
	#NACFC2016 – Kalydeco Seen to Improve Insulin Secretion in Patients with CF-related Diabetes. Read more &gt[semicolon]&gt[semicolon] http://bit.ly/2eMgaw6	#NACFC2016 – Kalydeco Seen to Improve Insulin Secretion in Patients with CF-related Diabetes. Read more &gt[semicolon]&gt[semicolon] http://bit.ly/2eMgaw6
	Breaking KemPharm Receives Clearance from FDA to Initiate Clinical Program for KP415, an Investigational Prodrug... http://tinyurl.com/jxm9m9a	KemPharm Receives Clearance from FDA to Initiate Clinical Program for KP415, an Investigational Prodrug of D-Thr.. http://www.financialbuzz.com/kempharm-receives-clearance-from-fda-to-initiate-clinical-program-for-kp-an-investigational-prodrug-of-d-threo-methylphenidate-for-the-treatment-of-adhd-580537
	Allergan teams with star paratriathlete in yet another effort to boost IBS-D med Viberzi: http://www.fiercepharma.com/marketing/allergan-teams-star-paratriathlete-yet-another-effort-to-boost-ibs-d-med-viberzi $AGN #pharma #marketing	Allergan teams with star paratriathlete in yet another effort to boost IBS-D med Viberzi http://www.pharmacommons.org/allergan/allergan-teams-with-star-paratriathlete-in-yet-another-effort-to-boost-ibs-d-med-viberzi/
FP	Tysabri time, at least I have a nice view (@ Cleveland Regional Medical Center in Shelby, NC)… https://twitter.com/i/web/status/790928103463915520	Tysabri time! #4 (@ Cleveland Regional Medical Center in Shelby, NC) https://www.swarmapp.com/c/lv33NyTQ2hs
	Don't forget your daytrana. You're throbbing without it. But wrath never goes away[semicolon] happiness is ever out of reach.	Don't forget your daytrana. You're capitalist without it. But distraction never goes away[semicolon] luck is ever out of reach.
	Don't forget your daytrana. You're capitalist without it. But distraction never goes away[semicolon] luck is ever out of reach.	Don't forget your daytrana. You're blunt without it. But consciousness never goes away[semicolon] commitment is ever out of reach.

Epidemico dataset with threshold at 19		
TP	Lower stomach pains. Idk if its from my essure procedure check up. Any advice #stayathomemom #essureproblems	Having pains on my lower stomach, idk if its from my essure procedure. Any advice #stayathomemom
	RT @clarejsurette: Still don't get how Humira hurts more than getting a tattoo did ?? #crohnie	Still don't get how Humira hurts more than getting a tattoo did ?? #crohnie
	#ESSURE ALMOST KILLED ME	@MommyinTX5 @BayerHealthCare We want answers #ESSURE ALMOST KILLED ME
	I eating shrooms popping OxyContin sipping lean straight, straight got me nodding	Im eating shrooms Poppin oxycodone Kissin on this lean straight Straight got me knoddin ??????????????
	RT @ericad4205: #essure stole my life for for 4 years and at the age of 28 I had to undergo a total hysterectomy #bayer #stopessure http://…	#essure stole my life for for 4 years and at the age of 28 I had to undergo a total hysterectomy #bayer #stopessure https://twitter.com/ericad4205/status/435409207827775490/photo/1
FP	Lumps after Juvederm in lips https://www.realself.com/question/south-carrollton-ky-lumps-upper-lip-after-juviderm-injections#1662942	Lumps after Juvederm https://www.realself.com/question/miami-fl-uneven-lumpy-lips-after-1ml-juvederm-visible-smile-stretch-lips#1998366
	I got a flu shot earlier and now I can't move my arm. ??	So, I got a Flu shot and now can't even move my arm.
	Uneven lips after Juvederm https://www.realself.com/question/los-angeles-ca-juvederm-liips-year-post-touch-uneven-lips#1412388	Uneven lips after Restylane injection https://www.realself.com/question/hampden-sydney-va-i-restylane-injection-lips-wednesday-bottom-lip-uneven#1648932 #fb
	Swelling after Restylane https://www.realself.com/question/tokyo-jp-normal-swelling-nose-and-lips#1549535	Swelling and Heat after Restylane Injections https://www.realself.com/question/chatsworth-ca-i-restylane-injected-face-feels-hot-swollen-allergic-reaction#1671476
	This lortab kicking my ass	this accutane is kicking my ass

Epidemico dataset with threshold at 50		
TP	"@LcplRoberts: Small pox vaccine in the left arm. Anthrax in the right. ????" don't forget the JEV vaccine ??	RT @LcplRoberts: Small pox vaccine in the left arm. Anthrax in the right. ????
	"@whoishegotyou: This Bacterial Meningococcal Vaccination got me feeling some type of way." ??????????	This Bacterial Meningococcal Vaccination got me feeling some type of way.
	Wtf my skin is drying out, I'm looking ashy. Time to stop using lubriderm and use Palmers cocoa butter (stereotyping myself lol)	My skin is drying out, I look do ashy. I need to stop using this lubriderm lotion and get cocoa butter (how stereotypical of me) lol
	@NicLJohnstx @mummy_bloggers I had gas&air,didn't work much, asked for pethidine,but was given too late as it didn't work either #mblogchat	@mummy_bloggers gas and air, tried water birth, but changed my mind, asked for pethidine, but it was given too late, no worked #mblogchat
	RT @BullDoza55: Never finna take #Concerta again I can't fucking sleep.....	Never finna take #Concerta again I can't fucking sleep.....
FP	Olanzapine made me fat Risperidone gave me headaches Quetiapine didn't do anything	Antipsychotic experiences Olanzapine made me fat Risperidone gave me headaches Aripiprazol fucked with my mood Quetiapine didn't stop voices
	To Much Weed And Codeine Fucked Up My Attitude	To much weed and codeine fuck up my attitude
	An hour after I take concerta I get super affectionate like I love everyone and everything	usually about an hour and a half after I take my concerta I get super affectionate about everything
	That HPV shot felt like a huge nail was stabbed into my arm??????	That HPV shot hurt felt like someone stabbed a huge nail into me ??????
	@NYNCpamiam @Debdebbailey @AlloccaMirella @THuntress17 @DeniseMira1 @TwilightGirl468 Yeah I got really bad from the flumist but now I am too	@Debdebbailey @AlloccaMirella @THuntress17 @DeniseMira1 @TwilightGirl468 @NYNCpamiam They put it on my chart no flumist bc the one year they

Discussion

In this study, probabilistic record linkage has been proven to be useful in the detection of duplicate posts in Twitter. Even though vigiMatch by design is conservative, i.e. aimed at high precision, about 17% of the posts in a data extract from Twitter could be detected as duplicates. Very few, about 1%, of the detected duplicate pairs were not true duplicates. Evaluating vigiMatch on the dataset provided by Epidemico we could find that, even though the dataset had been cleaned from duplicates using another method, an additional 9% of the posts should have been treated as duplicates, suggesting that probabilistic record linkage methods such as vigiMatch can provide further improvements over more traditional methods based on rules and/or Bloom filters.

Nonetheless, the estimated performance of our algorithm on the set of tweets we collected needs to be taken with a grain of salt. Since the data is unannotated from the start, we had to rely on the pairs we annotated during the active learning procedure. In each iteration, 20 pairs were sampled in each normalized score bin, however, the total number of pairs in the bins varied greatly (see Figure 2 for the distribution of scores) so the samples have various degrees of representativeness of their corresponding bins. It is not unlikely to miss duplicate pairs when sampling 20 out of 30 million pairs where the incidence of duplicate pairs is 5% (in fact, the probability of never sampling a duplicate pair in such case is roughly 37%), whereas we would sample the single duplicate pair in a bin of size 20. Considering the proportion of duplicate pairs to be zero instead of a small number in bins of large size can lead to a great overestimation of the recall. It is thus very possible that our recall estimates are upwardly biased. Better estimates of the precision can be obtained by directly sampling and manually reviewing the positive results.

The performance evaluation of our duplicate detection algorithm on the Epidemico dataset has been restricted to pairs containing at least one post annotated as a proto-AE post, to limit the number of pairs to manually review but still focus on what would be relevant for further analysis. Performance is likely to be higher in the entire dataset, as we observed a higher incidence of duplicates in the set of non proto-AE posts.

In its current implementation, our method presents one major limitation. The cut-off threshold to classify pairs of posts as duplicates cannot be pre-trained on a training dataset and then exported to a new dataset. Indeed, vigiMatch scores are computed using the word frequencies, which are dataset

dependent, hence the scores themselves are dataset dependent. Evaluating the precision of the algorithm at different thresholds by randomly sample and evaluate positive pairs is necessary to find the threshold that is appropriate for the data at hand and the desired performance. For unannotated data, recall will be difficult to evaluate. Other possibilities, not yet investigated, is to generalize the scores across datasets could include normalizing scores between datasets or normalizing the weights given for hits and misses in the model.

Changing the blocking design could also provide an improvement to our algorithm, in computational time. Currently, by comparing pairs based on all words they share instead of making all possible comparisons leads to a 23% reduction in the number of comparisons to be made. We could further improve this reduction by comparing pairs based on rarer words in common, relying on the assumption that core topics are conveyed via words of lower frequency and duplicates should share these rare words. Nonetheless, our algorithm already performs well in its current implementation: it can process 74 million pairs of posts in about 5 hours.

Apart from duplicate detection, our algorithm can be used to identify similar posts, which then could be used as help for annotators, to find out how other posts of similar content have been annotated previously and speed up the annotation process. In addition to a gain in time, such annotation help could lead to a better coherence in the choice of annotated terms. In the evaluation of the algorithm on the Epidemico dataset, we have seen several instances of proto-AE posts of similar nature that were either annotated differently or one of the two posts has been missed as a proto-AE post. At lower thresholds of scores, the algorithm could be used for topic clustering.

References

1. WHO. International Drug Monitoring: The Role of National Centres. 1972.
2. Inman WH. Postmarketing surveillance of adverse drug reactions in general practice. I: search for new methods. Br Med J (Clin Res Ed). 1981;282(6270):1131-2.
3. Twitter API [cited 2016 1 Sep]. Available from: https://api.twitter.com.
4. Ellenius J, Bergvall T, Dasgupta N, Hedfors S, Pierce C, and Norén GN. Medication Name Entity Recognition in Tweets Using Global Dictionary Lookup and Word Sense Disambiguation. Pharmacoepidemiol Drug Saf. 2016;25(Supplement S3):414.
5. Norén GN, Orre R, Bate A, Edwards IR. Duplicate detection in adverse drug reaction surveillance. Data Mining and Knowledge Discovery. 2007;14:305-28.
6. Copas JB, Hilton FJ. Record linkage: statistical models for matching computer records. J R Stat Soc Ser A Stat Soc. 1990;153(3):287-320.
7. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, Van Le H, et al. Evaluation of Facebook and Twitter Monitoring to Detect Safety Signals for Medical Products: An Analysis of Recent FDA Safety Alerts. Drug Saf. 2017;40(4):317-31.

Appendix

TermName	NumTweets	NumDuplicates
Denosumab	244	66
Seebri	3	2
Methylphenidate	477	132
Dapagliflozin	147	78
Gardasil	3172	792
Kalydeco	234	67
risperidone	719	158
daytrana	15	6
desloratadina	5	1
glycopyrronium	13	7
glycopyrrolate	34	7
risset	423	49
chymotrypsin	49	2
cipro	1040	90
natalizumab	151	91
zolpidem	551	35
cetirizine	472	97
novorapid	54	8
risperidona	5	2
zoledronate	32	14
desloratadine	44	8
insulin aspart	125	71
prolia	63	12
cervarix	235	33
rituxan	225	64
rituximab	878	295
ciplox	94	2
gardasil 9	56	11
glycopyrronium bromide	32	31
novolog	183	22
ciprofloxacin	390	116
ciprofloxacina	15	2
collagenase clostridium histolyticum	30	13
risperdal	753	155
varenicline	226	38
xiaflex	45	17
collagenase	80	31
mabthera	47	27
reclast	14	8
viberzi	139	47
zoledronic	24	3

aerius	161	56
clarinex	234	15
ivacaftor	195	92
tysabri	604	184
xgeva	54	24
zometa	419	26
santyl	12	8
zoledronic acid	129	47

SubstanceName	NumTweets	NumDuplicates
HPV vaccine	3375	819
Dapagliflozin	147	78
Ivacaftor	429	159
Glycopyrronium	96	47
Denosumab	365	102
Collagenase	172	69
Chymotrypsin	49	2
Rituximab	1150	386
Risperidone	1943	364
Cetirizine	474	97
Desloratadine	447	80
Insulin aspart	362	101
Ciprofloxacin	1565	210
Collagenase clostridium histolyticum	79	30
Varenicline	226	38
Ebastine	161	56
Eluxadoline	139	47
Loratadine;Pseudoephedrine	234	15
Zoledronic acid	624	98
Cyproheptadine	1040	90
Natalizumab	755	275
Zolpidem	551	35
Methylphenidate	499	138