Mining Archive.org's Twitter Stream Grab for Pharmacovigilance Research Gold

Ramya Tekumalla, Javad Rafiei Asl, Juan M. Banda
Georgia State University, Atlanta, Georgia, USA
rtekumalla1@gsu.edu, jasl1@student.gsu.edu, jbanda@gsu.edu

Abstract

In the last few years, Twitter has become an important resource for the identification of Adverse Drug Reactions (ADRs), monitoring flu trends, and other pharmacovigilance and general research applications. Most researchers spend their time crawling Twitter, buying expensive pre-mined datasets, or tediously and slowly building datasets using the limited Twitter API. However, there are a large number of datasets that are publicly available to researchers that are underutilized or unused. In this work, we demonstrate how we mined over 9.4 billion Tweets from archive.org's Twitter stream grab using a drug-term dictionary and plenty of computing power. Knowing that not everything that shines is gold, we used pre-existing drug-related datasets to build machine learning models to filter our findings for relevance. In this work, we present our methodology and the 3,346,758 identified tweets for public use in future research.

Introduction

The World Health Organization (WHO) defined Pharmacovigilance as "the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem" (World Health Organization, 2006) . The aim of pharmacovigilance is to enhance patient care and safety in relation to the use of medicines; and to support public health programmes by providing reliable, balanced information for the effective assessment of the risk-benefit profile of medicines. Traditionally, clinical trials are employed to identify and assess the profile of medicines. However, since they have limited ability to detect all ADRs due to factors such as small sample sizes, relatively short duration, and the lack of diversity among study participants, post marketing surveillance is required (Sultana, Cutroneo, and Trifirò, 2013). The Food and Drug Administration (FDA) provides several post marketing surveillance programs like FDA Adverse Event Reporting System (FAERS), MedWatch to report events, however, under-reporting limits its effectiveness. A review of 37 studies established that more than 90% of ADRs are estimated to be under-reported (Hazell and Shakir, 2006). Social media platforms like Twitter and Facebook contain an abundance of text data that can be utilized for pharmacovigilance (O'Connor et al., 2014). Many studies presented satisfactory results by utilizing social media for pharmacovigilance and helped create a curated dataset for drug safety (Sarker et al., 2015). Several other studies presented a connection between drugs and addictive behavior among students using Twitter (Hanson et al., 2013a and 2013b). However, it is challenging to use Twitter due to limitations with service providers who can export only 50,000 tweets per day. Further, usage of Twitter's API or software to extract tweets is extremely time-consuming and economically unviable, especially for obtaining tweets relevant to a particular domain. Additionally, machine learning and deep learning models need exorbitant amounts of training data to train a model and not much data is available publicly for training.

In this context, data sharing (Longo and Drazen, 2016) is a novel idea for research parasites to scavenge available datasets and apply their methodologies. Recent studies (Pasolli et al., 2016; Saez-Rodriguez et al., 2016; Warren, 2016) prove that data sharing improves quality and strengthens research. The increase in collaborative efforts will provide an opportunity for researchers to continually enhance research ideas and avoid redundant research efforts (Emmert-Streib, Dehmer, and Yli-Harja, 2016; Greene et al., 2017). In the past, few studies offered several pharmacovigilance insights or created curated datasets for drug safety (Klein et al., 2017; Nikfarjam et al., 2015; Sarker and Gonzalez, 2017). However, this is the largest publicly available dataset for research use of drug chatter from Twitter. As part of this research, we scavenged a large publicly available dataset and procured data related to pharmacovigilance. In this paper, we present a data corpus of 3,346,758 carefully filtered tweets. The deliverables (Tekumalla, Rafiei Asl, and Banda, 2019) include filtered 3,346,758 tweet ids, code to download, and separate tweets from the Internet Archive (IA). Due to Twitter's terms of service, tweet text cannot be shared. Therefore, tweet ids are publicly made available using Zenodo (Tekumalla Rafiei Asl, and Banda, 2019). The whole methodology can be reproduced using the deliverables. The released dataset adheres with FAIR principles (Wilkinson et al., 2016) in the following ways: The dataset is Findable as it can be accessed with a persistent DOI (Digital Object Identifier) in Zenodo. The dataset is Accessible through the DOI. The dataset contains only tweet identifiers as tweet text cannot be shared as per Twitter's terms of Service. However, tweets might be deleted either by Twitter or the user. In such cases, we can share the data on request while adhering to the Twitter data sharing policy. The tweet identifier can be hydrated to a tweet json object using tools like Social Media Toolkit (Tekumalla and Banda, 2020) or Twarc (twarc, n.d.). The hydrated tweets are json objects which are derived from JavaScript object notation syntax. JSON is a universally accepted format thus supporting Interoperability. This dataset is released with Creative Commons Attribution 4.0 International for Reusability.

Data Preparation

The Internet Archive (IA) (Machine, 2015) is a non-profit organization that builds digital libraries of Internet sites and other cultural artifacts in digital form and provides free access to researchers, historians, and scholars. The archive is mined from the Twitter stream API which according to Twitter is 1% sample of their daily tweets. This research utilizes the largest publicly available Twitter dataset in the Internet Archive, which contains several json files of tweets in tar files sorted by date for each month of the year. The tar file must be downloaded and decompressed before usage. A total of 9,406,233,418 (9.4 billion) tweets for the years 2012 to 2018 are available in this dataset, we filtered this data using a drug terms dictionary to identify drug-specific tweets. The time taken to download, process, and filter these tweets was 132 days.

Drug Dictionary Creation

The UMLS (National Library of Medicine, 2009) is a large, multi-purpose and multilingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. The UMLS includes the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon and Lexical Tools. Metathesaurus, the biggest component of UMLS, was utilized in creating the drug dictionary, more specifically the RxNorm (National Library of Medicine, 2008) vocabulary. This vocabulary provides normalized names for clinical drugs and link names to the drug vocabulary commonly used in pharmacy management and drug interaction software. The MRCONSO table was filtered using RxNorm and Language of Term (LAT), which was set to English. The filtered table contained a total of 279,288 rows. Since the dictionary was used on Twitter data and the total number of characters allowed in a tweet was 140 (until 2017) and 280 (from October 2017 onwards), we eliminated all the strings of length less than or equal to 3 (too ambiguous) and greater than or equal to 100. This was due to a less likely chance for tweets to contain drug names that were as short as 3 characters or as long as 100 characters. Further, we removed strings such as "2,10,15,19,23-pentahydrosqualene" which are chemical compounds. This elimination was based on the premise that users would find it cumbersome and tedious to type detailed chemical names of drugs, especially on social media. Additionally, we removed 50 terms like "disk, foam, bar-soap, sledgehammer, cement, copper, sonata" as these terms are not commonly used as drug names and in pharmacovigilance. After deleting the common terms and chemical compounds, only 266,556 rows were available of which five term types were used in the drug dictionary for the research. The dictionary also consists of a Concept Unique Identifier (CUI) to which strings with the same meaning are linked. The CUI is used in order to ensure that the meanings are preserved over time regardless of the different terms that are used to express those meanings. All the strings have been converted to lowercase and trimmed of white spaces. A total of 111,518 unique strings were used in total to create the drug dictionary. Table 1 represents the number of strings used for each term type and Table 2 contains sample rows from the drug dictionary.

Term Type	Example	# Strings
Ingredients (IN)	Fluoxetine	11,427
Semantic Clinical Drug Component (SCDC)	Fluoxetine 4 MG/ML	27,038
Semantic Branded Drug Component (SBDC)	Fluoxetine 4 MG/ML [Prozac]	17,938
Semantic Clinical Drug (SCD)	Fluoxetine 4 MG/ML Oral Solution	35,112
Semantic Branded Drug (SBD)	Fluoxetine 4 MG/ML Oral Solution [Prozac]	20,003

Table 1: Term types, their definitions and Number of strings

Concept Unique Identifier (CUI)	Term String
C0290795	adderall
C0700899	benadryl
C0025219	melatonin
C0162373	prozac
C0699142	tylenol

Table 2: Sample Drug Dictionary

Methods

In order to identify drug-specific tweets that would be useful for pharmacovigilance, we applied the drug dictionary on the Internet Archive Twitter dataset. We filtered the dataset using spaCy, an open-source library in Python. We used the matcher in spaCy which would match sequences of tokens based on pattern rules. Subsequently, the program generates an output file with the filtered tweets if it finds a match with the drug dictionary in the tweet text. Tweets are retrieved only if their language is set to English and if they are not retweeted. Initially, the method was performed on 2018 data, with our results showing that the maximum number of tweets that got separated consisted of a single drug string (one term). We speculate that, since Twitter has a limitation on the number of characters, people tend to write abbreviated terms or single terms that are either drug names or ingredients, instead of a drug string that consists of 4 or 5 terms. A single string dictionary is created from the five dictionaries with a total of 13,226 unique single terms. A total of 6 programs are run on each month for the year 2018. Only 10 months of data was available for the year 2018. Number of tweets obtained for four months for the year 2018 when used on six dictionaries are presented in table 3. Note that since SCD, SBD and SBDC did not yield any matches in the 2018, we removed the from the table for simplicity and clarity.

Month	Total	Single string	Ing.	SCDC
Jan.	134,747,413	83,583	27,718	15
Aug.	141,132,076	67,227	21,335	13
Sept.	133,068,824	64,123	22,230	22
Oct.	132,297,280	68,221	22,955	17
Total	**1,102,507,263**	**385,503**	**196,788**	**112**

Table 3: Number of tweets obtained for each month from the Internet Archive Dataset in 2018.

The SCD, SBD and SBDC dictionaries did not yield any tweets from 2018. In order to determine the reason, we examined and analyzed the dictionaries. For each term type in the drug dictionary, we calculated the lengths of all drug strings and identified the number of characters at each length ranging between 4 and 99 characters. Further, we also noted the average and median lengths of the drug strings. Table 4 depicts detailed statistics for each term type.

SBD, SCD and SBDC had the highest number of lengthy drug strings. The following drug string from SCDC drug dictionary, "pneumococcal capsular polysaccharide type 33f vaccine 0.05 mg ml", has 64 characters. It is impractical to type the whole drug string in a tweet without an error. 90% of the tweets obtained from the SCDC were advertisements on either promoting the product or selling the product. Further examining all the tweets, we eliminated the 4 dictionaries (SCD, SCDC, SBD, SBDC) and used the single string and ingredients dictionary since it saves an enormous amount of computation time.

Term Type	Minimum length (#)	Maximum length (#)	Average length
Ingredients	4 (11)	99 (1)	21.556
SCDC	10 (16)	93 (2)	24.152
SBDC	19 (1)	99 (54)	43.757
SCD	20 (1)	99 (99)	48.531
SBD	32 (3)	99 (84)	57.818

Table 4: Statistics for each drug term type

We made the code publicly available for reproducibility. A total of 132 days were required to download, unzip, and filter the tweets using the drug dictionary. For all the years, each month was downloaded individually, unzipped to retrieve the json file and then the tweets were filtered using the drug dictionary. Typically, for a month, the method would require 10 minutes to download, 5 hours to unzip and 2 days to filter tweets on an IBM Blade Server with 768 GB RAM, 2 x Intel® Xeon® E5-269880 Processors, with 40 cores each, and 12TB of hard disk space. Table 5 represents the number of tweets retrieved for the years from 2012 to 2018 when used with the two remaining drug dictionaries.

The single string and ingredients dictionary was used on the IA dataset and a total of 6,703,331 (6.7 million) tweets were retrieved from 9,406,233,418 (9.4 billion) tweets. After eliminating duplicate tweets, a total of 6,703,166 were retrieved. We examined the retrieved tweets and found that more than 50% of the tweets are not relevant to pharmacovigilance. This is because some drug strings are used in common terminology and in other fields like math, technology etc. For example, the drug string "tablet" was used in reference to the electronic gadgets (Samsung, Microsoft tablets). In order to eliminate the tweets that are relevant to other domains and not pharmacovigilance, we employed machine learning and deep learning classification models to filter tweets.

Month	Total tweets	Single string	Ingredients
2018	1,102,507,263	588,854	26,034
2017	1,448,114,354	1,079,616	52,058
2016	1,427,468,805	1,252,057	62,514
2015	1,224,040,556	1,149,736	55,367
2014	1,086,859,898	873,937	47,953
2013	1,871,457,526	1,463,184	85,618
2012	1,245,785,016	76,795	4,139
Total tweets	**9,406,233,418**	**6,152,862**	**314,680**

Table 5: Number of single and ingredient tweets obtained from total tweets

Classification

Since the filtered tweets contain a number of irrelevant tweets, we experimented with several classical machine learning and deep learning models on the filtered tweets to clean the tweets. The tweets obtained after classification can be used for training different machine learning and deep learning models by other researchers. Since there are no trainable datasets that we could make use of, we created a dataset utilizing annotated datasets from different sources (Ginn et al., 2014; Klein et al., 2017; Nikfarjam et al., 2015; Sarker & Gonzalez, 2017). We emphasize that we did not annotate or create any annotated set of tweets ourselves.

Classical Models

We collected 259,042 tweets that only have drug strings from multiple papers on pharmacovigilance using social media (Ginn et al., 2014; Klein et al., 2017; Nikfarjam et al., 2015; Sarker & Gonzalez, 2017) and downloaded all the tweets available through them. These tweets were annotated by different annotators as part of their research. The collected 259,042 tweets from multiple pharmacovigilance papers were labelled as "drug" tweets. Additionally, we randomly collected 300,208 non-drug tweets from multiple years from the Internet Archive and labelled them as "non-drug" tweets. Pre-processing was performed on the downloaded tweets by removing links and emojis and only tweet text was separated. A total of 559,250 tweets were used as an annotated training set where only drug tweets were the actual annotated tweets collected from different sources. We experimented with five classifiers: Naive Bayes, Logistic Regression, Support Vector Machines (SVM), Random Forest and Decision Trees using the scikit-learn (Pedregosa et al., 2011). Support-Vector Machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. We used a LinearSVC which is similar to SVC, but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. We used the Multinomial Naive Bayes which implements the naive Bayes algorithm for multinomial distributed data and is one of the two classic naive Bayes variants used in text classification. A Random Forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The Decision Tree Classifier uses a CART algorithm (Classification And Regression Tree). CART is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. However, the scikit uses an optimized version of the CART which does not support categorical values.

Each classifier model is applied on the stratified 75-25% (training - test) split of the annotated training set. We calculated precision, recall, and F-score to evaluate each classifier and the results are tabulated in Table 6.

Classifier	Precision	Recall	F-measure	Accuracy
Logistic Regression	0.7535	0.7814	0.7672	0.8267
Naive Bayes	0.7106	0.8281	0.7649	0.8140
SVM	***0.7773***	***0.8091***	***0.7929***	***0.8456***
Decision Tree	0.7274	0.5120	0.6010	0.7516
Random Forest	0.7406	0.6814	0.7097	0.7963

Table 6: Classification metrics for the classical machine learning models

Deep Learning Models

For the classification task, we experimented with six deep learning techniques from the Matchzoo framework (Guo et al., 2019) : MVLSTM, DUET, KNRM, CONVKNRM, DSSM, and ARC-II (Dai et al., 2018; Huang et al., 2013; Hu et al., 2014; Mitra, Diaz, and Craswell, 2017; Wan et al., 2016; Xiong et al., 2017). DUET is applied for the document ranking task and is composed of two separate deep neural networks, one matches the query and the document using a local representation, and another matches the query and the document using learned distributed representations. KNRM is a kernel-based neural model for conducting the document ranking task by using three sequential steps: 1) a translation matrix to model word-level similarities using word embeddings. 2) a modern kernel-pooling technique to use kernels for multi-level soft match features extraction. 3) a learning-to-rank layer that combines those features into the final ranking score. CONVKNRM uses CNNs to compose n-gram embeddings, and cross-matches n-grams of various lengths. It applies kernel pooling to extract ranking features, and uses learning-to-rank to obtain the final ranking score. MVLSTM conducts sentence matching with multiple positional sentence representations where each representation is generated by a bidirectional LSTM. The final score is produced by aggregating interactions between these different positional sentence representations. ARC-II focuses on sentence matching by naturally combining the hierarchical sentence modeling through layer-by-layer composition and pooling and capturing of the rich matching patterns at different levels of abstraction. DSSM aims to rank a set of documents for a given query. First, a non-linear projection maps the query and documents to a common semantic space. Then, the relevance of a document with the query is calculated as cosine similarity between their vectors in the semantic space. These deep models are general-purpose models and can be used for different text matching tasks such as document retrieval, conversational response ranking, and paraphrase identification. Precision, recall, F-measure, and accuracy metrics are tabulated in Table 7.

Model	Precision	Recall	F-measure	Accuracy
ARC-II	**0.9679**	**0.9581**	**0.9630**	**0.9731**
DUET	0.9635	0.9579	0.9607	0.9715
DSSM	0.6457	0.3904	0.4866	0.7001
KNRM	0.9585	0.9158	0.9367	0.9549
MVLSTM	0.9781	0.9047	0.9400	0.9579
CONV-KNRM	0.9692	0.9402	0.9545	0.9673

Table 7: Classification metrics for the deep learning models

Calculating cutoff thresholds

We applied all classical and deep learning models on the filtered 6 million tweets to predict the probability score of each tweet from the Internet Archive dataset. In order to determine the most optimal probability cutoff, we applied mixture models concepts (Budczies et al., 2012). The way this methodology works is by taking all the probability scores and dividing them into several hundreds of bins. A histogram of probability frequency is determined by calculating the number of observations in each bin. Based on the hypothesis of mixture models, probability scores are distributed according to a mixture of two Gaussian distributions (drug and non-drug tweets). Finally, two highest peaks of two Gaussian distributions and one valley with most depth between the two peaks are detected and the valley's deepest point is used as a cut-off point (threshold). In Figure 1 and 2, we plot the number of tweets that have a given probability score. Starting from an assigned probability of one, we cumulatively count the number of tweets we would keep at any given probability threshold. These plots allow us to see the selectivity of each model and the number of tweets at each threshold limit. Note that the optimal cutoff threshold is displayed next to the model name in the figure legend.

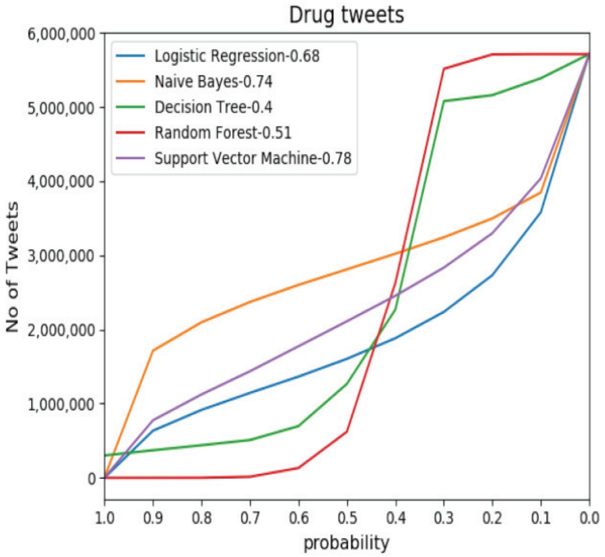

Figure 1. Probabilities of drug tweets using Classical Models

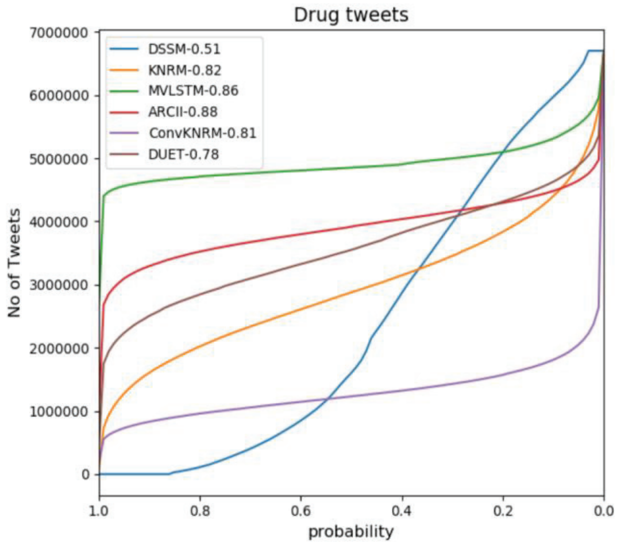

Figure 2. Probabilities of drug tweets using Deep Learning Models

Based on Tables 6 and 7, and the cutoff Figures 1 and 2, we selected ARC-II as the model to use to classify the relevance of the tweets. This deep learning model performed the best in terms of F-measure and accuracy, two of the metrics we deemed most relevant to identify useful tweets. The trained models can be shared upon request. After the classification filtering, we examined all the retrieved tweets and calculated the drug occurrences. We identified 6,867 unique drug strings in 3,346,758 million tweets. At the moment, this is the largest publicly available dataset for research use of drug chatter from Twitter. Please note that the released dataset consists of only tweet identifiers. The tweet identifiers can be hydrated to a json object using either Social Media Mining Toolkit (Tekumalla and Banda, 2020) or twarc (twarc, n.d.). The entire methodology of the research is depicted in Figure 3 and Figure 4 depicts the popular drug strings and the number of occurrences for each drug string in the classified tweets.

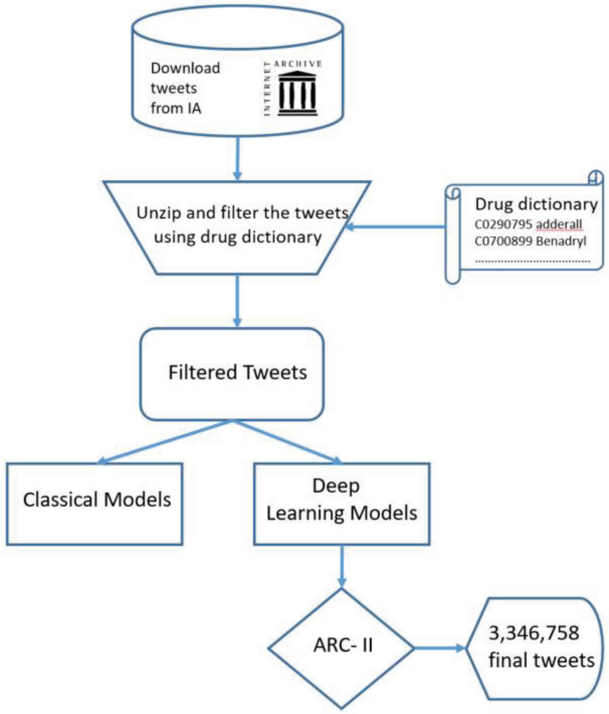

Figure 3. Methodology of Mining archive.org's Twitter stream grab for Pharmacovigilance research gold

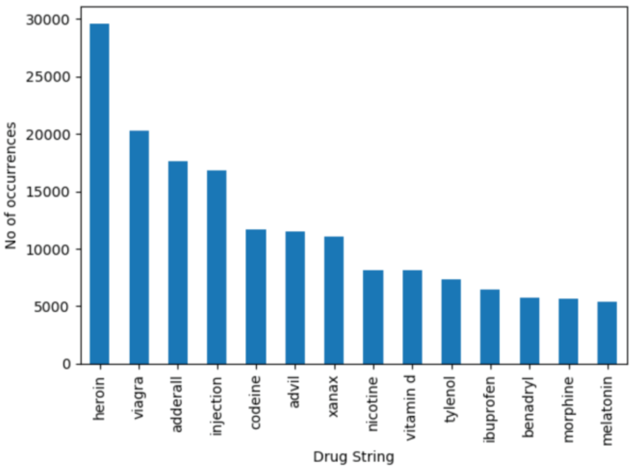

Figure 4. Popular drug string occurrences after filtering and classification of tweets

Cocaine was the most popular drug string with 59,397 occurrences, however, we eliminated it from the plot since it was used as a recreational drug than a medical drug. The following are a few examples of the tweets. The drug strings are highlighted in bold.

1. “my head still hurting tho . my mama gave me some **ibuprofen** .”
2. “im so stupid for taking **benadryl** in the morning #sleeeeeepy”
3. “having to stop the **vicodin** completely because its making me sick seriously anything . #letthepain-begin”.
4. “i got very sick from **effexor** . also lost 2 years but ive recovered fully and life’s better than ever . hang in there”.
5. i took some **tylenol** with **codeine**. im sleepy , but i have to change the gauze in my mouth because it wont stop bleeding.”

Future Work

The proposition of this paper is to utilize publicly available resources and employ Machine and Deep Learning techniques to create a dataset that can be made available for pharmacovigilance research. We believe that we cannot train models with a very limited amount of manually annotated tweets, but we can use the theory of noisy labeling to create more robust models with silver standards (Agarwal et al., 2016; Han et al., 2019; Paul et al., 2019). However, there are a few limitations, which we would like to address in our future work. This research utilizes only English tweets since there were no publicly available annotated drug tweets for other languages. Currently, validation is performed only on the classification model but not the annotated dataset. Furthermore, the annotated drug tweets used in the training data were collected from publicly available sources and are labelled as drug tweets. Hence, edge cases such as ambiguous tweets were not considered. The language in Twitter is neither professional nor standard. Therefore, there would be a great possibility of spelling errors and slang. In the future, we will employ a spelling correction module, which includes the tweets with incorrect drug spellings, which greatly improves the scale of the dataset. Further, we would like to develop an improved annotated dataset, which can be utilized as a gold standard dataset, following a tri-modal distribution of probabilities where the edge cases are considered.

Conclusion

In this paper, we scavenged a publicly available Twitter dataset, Internet Archive, mining over 9.4 billion tweets. Using a simple drug dictionary and plenty of computing power, we filtered 6 million tweets with relevant drug terms in them. In order to determine the viability of the filtered tweets for research work, we used publicly available, manually and expertly curated tweet datasets to build classification models to identify the relevant (or similar) tweets in our dataset. Overall, these tasks took around 150 days for downloading, filtering and classification, in order to retrieve 3,346,758 tweets, which can be utilized for drug safety research and as a training set for other supervised methods by researchers. Further, we believe that this approach can be reused and extended to several other domains by changing the dictionaries and the filtering mechanisms.

References

Agarwal, V.; Podchiyska, T.; Banda, J. M.; Goel, V.; Leung, T. I.; Minty, E. P.; Sweeney, T. E.; Gyang, E.; and Shah, N. H. 2016. Learning statistical models of phenotypes using noisy labeled training data. Journal of the American Medical Informatics Association: JAMIA, 23(6), 1166–1173. https://doi.org/10.1093/jamia/ocw028

Budczies, J.; Klauschen, F.; Sinn, B. V.; Győrffy, B.; Schmitt, W. D.; Darb-Esfahani, S.; and Denkert, C. 2012. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. PloS One, 7(12), e51862. https://doi.org/10.1371/journal.pone.0051862

Dai, Z.; Xiong, C.; Callan, J.; and Liu, Z. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. Conference on Web Search and Data …. https://dl.acm.org/citation.cfm?id=3159659

Emmert-Streib, F.; Dehmer, M.; and Yli-Harja, O. 2016. Against Dataism and for Data Sharing of Big Biomedical and Clinical Data with Research Parasites. Frontiers in Genetics, 7, 154. https://doi.org/10.3389/fgene.2016.00154

Ginn, R.; Pimpalkhute, P.; Nikfarjam, A.; and Patki, A. 2014. Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark. Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.672.9123andrep=rep1andtype=pdf

Greene, C. S.; Garmire, L. X.; Gilbert, J. A.; Ritchie, M. D.; and Hunter, L. E. 2017. Celebrating parasites [Review of Celebrating parasites]. Nature Genetics, 49(4), 483–484. https://doi.org/10.1038/ng.3830

Guo, J.; Fan, Y.; Ji, X.; and Cheng, X. 2019. MatchZoo: A Learning, Practicing, and Developing System for Neural Text Matching. In arXiv [cs.IR]. arXiv. http://arxiv.org/abs/1905.10289

Han, J.; Luo, P.; and Wang, X. 2019. Deep self-learning from noisy labels. Proceedings of the IEEE International Conference on Computer Vision, 5138–5147. http://openacess.thecvf.com/content_ICCV_2019/html/Han_Deep_Self-Learning_From_Noisy_Labels_ICCV_2019_paper.html

Hanson, C. L.; Burton, S. H.; Giraud-Carrier, C.; West, J. H.; Barnes, M. D.; and Hansen, B. 2013a. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. Journal of Medical Internet Research, 15(4), e62. https://doi.org/10.2196/jmir.2503

Hanson, C. L.; Cannon, B.; Burton, S.; and Giraud-Carrier, C. 2013b. An exploration of social circles and prescription drug abuse through Twitter. Journal of Medical Internet Research, 15(9), e189. https://doi.org/10.2196/jmir.2741

Hazell, L. and Shakir, S. A. W. 2006. Under-reporting of adverse drug reactions : a systematic review. Drug Safety: An International Journal of Medical Toxicology and Drug Experience, 29(5), 385–396. https://doi.org/10.2165/00002018-200629050-00003

Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. Proceedings of the 22Nd ACM International Conference on Information and Knowledge Management, 2333–2338. https://doi.org/10.1145/2505515.2505665

Hu, B.; Lu, Z.; Li, H.; and Chen, Q. 2014. Convolutional Neural Network Architectures for Matching Natural Language Sentences. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27 (pp. 2042–2050). Curran Associates, Inc. http://papers.nips.cc/paper/5550-convolutional-neural-network-architectures-for-matching-natural-language-sentences.pdf

Klein, A.; Sarker, A.; Rouhizadeh, M.; O'Connor, K.; and Gonzalez, G. 2017. Detecting personal medication intake in Twitter: an annotated corpus and baseline classification system. BioNLP 2017, 136–142. https://www.aclweb.org/anthology/papers/W/W17/W17-2316/

Longo, D. L., and Drazen, J. M. 2016. Data Sharing. The New England Journal of Medicine, 374(3), 276–277. https://doi.org/10.1056/NEJMe1516564

Machine, W. 2015. The Internet Archive. Searched for Http://www.Icann.Org/icp/icp-1.Htm. https://rena.mpdl.mpg.de/rena/Record/ERS000001789

Mitra, B.; Diaz, F.; and Craswell, N. 2017. Learning to Match Using Local and Distributed Representations of Text for Web Search. Proceedings of the 26th International Conference on World Wide Web, 1291–1299. https://doi.org/10.1145/3038912.3052579

National Library of Medicine. 2008, October 8. RxNorm [Internet]. National Library of Medicine (US). http://www.nlm.nih.gov/research/umls/rxnorm/

National Library of Medicine (US). 2009, September. UMLS® Reference Manual [Internet] : 2, Metathesaurus. National Library of Medicine (US). https://www.ncbi.nlm.nih.gov/books/NBK9684/

Nikfarjam, A.; Sarker, A.; O'Connor, K.; Ginn, R.; and Gonzalez, G. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association: JAMIA, 22(3), 671–681. https://doi.org/10.1093/jamia/ocu041

O'Connor, K.; Pimpalkhute, P.; Nikfarjam, A.; Ginn, R.; Smith, K. L.; and Gonzalez, G. 2014. Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium, 2014, 924–933. https://www.ncbi.nlm.nih.gov/pubmed/25954400

Pasolli, E.; Truong, D. T.; Malik, F.; Waldron, L.; and Segata, N. 2016. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Computational Biology, 12(7), e1004977. https://doi.org/10.1371/journal.pcbi.1004977

Paul, D.; Singh, M.; Hedderich, M. A.; and Klakow, D. 2019. Handling Noisy Labels for Robustly Learning from Self-Training Data for Low-Resource Sequence Labeling. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1903.12008

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, É. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research: JMLR, 12(Oct), 2825–2830. http://www.jmlr.org/papers/v12/pedregosa11

Ramya Tekumalla, Javad Rafiei Asl, Juan M. Banda. n.d.. InternetArchive-Pharmacovigilance-Tweets. Github. Retrieved August 14, 2019, from https://github.com/thepanacealab/InternetArchive-Pharmacovigilance-Tweets

Saez-Rodriguez, J.; Costello, J. C.; Friend, S. H.; Kellen, M. R.; Mangravite, L.; Meyer, P.; Norman, T.; and Stolovitzky, G. 2016. Crowdsourcing biomedical research: leveraging communities as innovation engines. Nature Reviews. Genetics, 17(8), 470–486. https://doi.org/10.1038/nrg.2016.69

Sarker, A.; Ginn, R.; Nikfarjam, A.; O'Connor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; and Gonzalez, G. 2015. Utilizing social media data for pharmacovigilance: A review. Journal of Biomedical Informatics, 54, 202–212. https://doi.org/10.1016/j.jbi.2015.02.004

Sarker, A.; and Gonzalez, G. 2017. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. Data in Brief, 10, 122–131. https://doi.org/10.1016/j.dib.2016.11.056

Sultana, J.; Cutroneo, P.; and Trifirò, G. 2013. Clinical and economic burden of adverse drug reactions. Journal of Pharmacology and Pharmacotherapeutics, 4(Suppl 1), S73–S77. https://doi.org/10.4103/0976-500X.120957

Tekumalla, R. and Banda, J. 2020. Social Media Mining Toolkit (SMMT). Under review. Genomics and Informatics. https://github.com/thepanacealab/SMMT

Tekumalla, R.; Rafiei Asl, J.; and Banda, J. 2019. Mining Archive.org's Twitter Stream Grab for Pharmacovigilance Research Gold Dataset [Data set]. https://doi.org/10.5281/zenodo.3571328

twarc. (n.d.). Github. Retrieved March 23, 2020, from https://github.com/DocNow/twarc

Wan, S.; Lan, Y.; Guo, J.; Xu, J.; Pang, L.; and Cheng, X. 2016. A deep architecture for semantic matching with multiple positional sentence representations. Thirtieth AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/11897

Warren, E. 2016. Strengthening Research through Data Sharing. The New England Journal of Medicine, 375(5), 401–403. https://doi.org/10.1056/NEJMp1607282

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R,; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. L.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, R.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, L.; Zhao, J.; and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

World Health Organization, The safety of medicines in public health programmes: Pharmacovigilance an essential tool. 2006. http://www.who.int/medicines/areas/quality_safety/safety_efficacy/

Xiong, C.; Dai, Z.; Callan, J.; Liu, Z.; and Power, R. 2017. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 55–64. https://doi.org/10.1145/3077136.3080809