Construction of a Personal Experience Tweet Corpus for Health Surveillance

Keyuan Jiang **Ricardo A. Calix** **Matrika Gupta**
Department of Computer Information Technology & Graphics
Purdue University Northwest
{kjiang,rcalix,gupta297}@pnw.edu

Abstract

Studies have shown that Twitter can be used for health surveillance, and personal experience tweets (PETs) are an important source of information for health surveillance. To mine Twitter data requires a relatively balanced corpus and it is challenging to construct such a corpus due to the labor-intensive annotation tasks of large data sets. We developed a bootstrap method of finding PETs with the use of the machine learning-based filter. Through a few iterations, our approach can efficiently improve the balance of two class dataset with a reduced amount of annotation work. To demonstrate the usefulness of our method, a PET corpus related to effects caused by 4 dietary supplements was constructed. In 3 iterations, a corpus of 8,770 tweets was obtained from 108,528 tweets collected, and the imbalance of two classes was significantly reduced from 1:31 to 1:3. In addition, two out of three classifiers used showed improved performance over iterations. It is conceivable that our approach can be applied to various other health surveillance studies that use machine learning-based classifications of imbalanced Twitter data.

1 Introduction

As defined by the Merriam-Webster Dictionary, surveillance is the act of carefully watching someone or something. In the health field, the WHO defines that public health surveillance is the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice. Information directly reported by patients is of significant importance, and having an efficient way of obtaining and analyzing this data is very important. Because of mobile phones and other technologies, patients are inclined to post information on the web. This represents a great opportunity for those concerned with health surveillance if they can only mine the data. As such, the critical issue is where and how to obtain and analyze this health surveillance data.

Nowadays, social media has become a natural platform through which people communicate and share their thoughts, opinions, and experiences. Topics of communication span to a broad range from politics to entertainment to hobbies. Many people are also willing to discuss their personal experiences related to their health problems and treatments on social media. Studies have shown that general purpose social media such as Twitter can be used for surveillance of health-related issues (Dredze, 2012). Examples include: affluenza pandemics (Chew and Eysenbach, 2010; Signorini et al., 2011; Collier et al., 2011; Bilge et al., 2012; Nagel et al., 2013; Gesualdo et al., 2013; Broniatowski et al., 2013; Fung et al., 2013; Nagar et al., 2014), Haitian cholera outbreak (Chunara et al., 2012), Ebola outbreak (Odlum and Yoon, 2015), nonmedical use of a psychostimulant drug (Adderall) (Hanson et al., 2013), drug abuse (Chary et al., 2013), smoking (Sofean and Smith, 2012), suicide risks (Jashinsky et al., 2014), migraine headaches (Nascimento et al., 2014), pharmaceutical product safety (Freifeld et al., 2014; Coloma et al., 2015; Jiang and Zheng, 2013; Sarker et al., 2015), disease outbreaks during festivals (Yom-Tov et al., 2014), detection of Schizophrenia (McManus et al., 2015), foodborne illness (Harris et al., 2014), and even dental pains (Heaivilin et al., 2011).

A common challenge identified in these types of studies is the difficulty in separating the useful

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 128–135,
Berlin, Germany, August 12, 2016.

or "on-topic" tweets from the majority of the irrelevant tweets. This poses the challenge of finding the tweets that can help to perform the heath surveillance tasks while ignoring the rest.

Twitter is a micro blogging platform on which messages of up to 140 characters can be posted. Despite the shortness of the messages, the size of Twitter user pool may still mean that a lot of information can be posted. As such, for any given topic, there may be a good number of on-topic tweets and a much larger set of off-topic tweets. As a result, one of the key questions to address is how to obtain the relevant data.

In this study, the term personal experience tweet (PET) is used to describe the tweets that are relevant to the analysis. PETs, therefore, are tweets that describe a person's encounters, observations, and important events related to his or her life. In the case of health surveillance, such experience can be related to changes of a person's health, an illness, a disease, or a treatment. In other words, if any of the above affects an individual it signifies a personal experience. For example, if a medicine causes a person to vomit or improves the person's sleeping behavior, then the person is said to have some experience with the medicine. Personal experience tweets (PETs) are an important source of information for health surveillance using Twitter data.

Given the sheer volume of daily posts, Twitter data are known to contain a significant amount of irrelevant off-topic posts (e.g. news, sales promotions, spam, etc.). This can easily result in collections of Twitter data with a significant bias toward the irrelevant posts. For example, in a study of 2 billion tweets collected from May 2009 to October 2010, Bian and colleagues (Bian et al., 2012) found only 489 on-topic tweets for the 5 medicines being studied in clinical trials. As can be seen, from this study discovering on-topic tweets can be a challenge in research problem. Given all the previously stated issues, obtaining relevant data and constructing a relatively balanced corpus can be challenging and a good collection process must be implemented. This paper will discuss the data collection process, the automatic filtering approach, the annotation, and results of the analysis of the corpus. Issues related to class imbalance are also discussed.

Specifically, this study addresses the following research questions: (1) can an automated filtering algorithm help to speed up manual annotation of a PET corpus and (2) can the automated filtering approach help to address the class imbalance issues inherent in Twitter data?

2 Related Work

There have been many studies that validate the use of general purpose social media such as Twitter for surveillance of health related issues. Many of these surveillance activities involve using the information reported by the patients who share their personal health experience on social media. Efforts have been made to construct health-related Twitter corpora (Paul and Dredze, 2012; Collier et al., 2011; Ginn et al., 2014).

Using Mechanical Turk, Dredze's group (Paul and Dredze, 2012) created a corpus of 5,128 tweets classified as related to health or unrelated to health. The results showed only 36.1% of the labeled tweets were health related. It is unclear how the tweets were selected into the corpus.

Collier and colleagues (Collier et al., 2011) created a 5,283 tweets corpus related to influenza from 225,000 tweets collected from March 2010 to April 30th, 2010. These tweets in 5 classes were selected using *hand built patterns* which were unexplained by the authors, and annotated by a single annotator. For each of the 5 classes, the ratio of negative tweets to positive tweets was 2.52, 1.16, 1.95, 7.19 and 2.53 respectively, indicating that there were more negative tweets than positive ones in each class.

In studying adverse drug reactions from Twitter data, Ginn et al. (Ginn et al., 2014) collected 187,450 tweets over 6 months with 74 carefully selected drug names. 71,571 tweets were retained after removing those containing URLs, which were considered as advertisements. Out of 71,571 tweets, 10,822 were randomly chosen with a cap of 300-500 per drug. The 10,822 tweets were manually annotated by three annotators. Among 10,822 tweets, only 1,200 (11%) tweets contain adverse drug reactions (ADRs), showing the imbalance ratio of 1:8. The authors also reported a Kappa inter-annotator agreement metric with a value of 0.69.

3 Methodology

The purpose of health surveillance is to monitor the status of health conditions. To track health information using Twitter data, a data set of Twit-

ter texts is needed. With this dataset, a methodology can be devised to identify the effects in the text. The challenge is in discovering the relevant tweets. Our initial inspection of tweets collected using 4 dietary supplement names showed that many of the tweets were not personal experience tweets relevant to the work. Manual annotation is an expensive process, especially when using large datasets which contain very few on-topic samples. Therefore, an automated filtering tool was needed to address these issues. One of the purposes of this study is to speed up the process of annotation. Many studies have used manual or rule-based approaches for annotation. However, these approaches are time consuming. In this paper, a machine learning-based approach is proposed to try to filter out off-topic tweets.

Inspired by the bootstrap method, we developed an iterative approach of creating Twitter corpus. It starts with a small set of annotated tweets (seed). In each iteration, the annotated tweets (in the training set which is the corpus) are used to retrain classifiers, and the predicted tweets of PET class from the trained classifiers are annotated and added to the training set, in an attempt to obtain a less imbalanced corpus.

In this section, we present our method of finding personal experience tweets and its application in constructing a PET corpus related to the effects caused by 4 dietary supplements. An automated filter was used to try to remove irrelevant samples before the data set was given to annotators. A description of the creation of the PET corpus using this filter is also presented and discussed. The next few sections of the paper describe in more detail the various considerations and methodology used to create the corpus.

3.1 Corpus Construction Procedure

This study was done with the help of two annotators, who were graduate students majoring in biology and computer information technology. They independently labeled the same tweets with personal experience tags if they contained the name of any supplements and stated the experiencing of using the supplements. Below are examples of PET tweets.

Example 1:

1. *melatonin gives me some messed up dreams.. or i just have awful dreams and melatonin makes me remember them. either way i dont like it.*

Example 2:

2. *look into St. John's Wort. Actually helps calm me down at night to sleep. always had the same issues.*

First, a small number of tweets were randomly selected as a training set and were annotated manually by annotators. This was a single non-repetitive step to create a seed set. Next, three classifiers were trained using this training set and then used to classify a test set with more tweets, yielding a PET set and a non-PET set. The PET set was then labeled by annotators, and annotated tweets were added to the training set (corpus). Classifiers were retrained with the updated corpus and then used on a new batch of test data. These steps repeated until a relatively balanced corpus was achieved.

Although investigating and annotating only the predicted PET class significantly reduce the effort needed for annotation, it could potentially introduce bias undermining the representation of non-PET (majority) tweets. To compensate this potential bias, we intentionally added a small number of non-PET tweets to the training set in each iteration (Step 06 below).

The above steps are summarized in the following algorithm.

Algorithm **ConstructTweetCorpus()**
Input: A set of tweets **T**, balance ratio β,
accuracy δ
Output: A tweet corpus *T*
01: Randomly choose a small collection of n tweets from **T** as a training set denoted by *T*
02: Annotate *T*
03: Train classifiers with *T*
04: **Do while** balance ratio of $T < \beta$ and/or accuracy of classifiers $< \delta$
05: Select a collection of *l* new tweets from **T** as test set denoted by T_l
06: Classify T_l using trained classifiers, yielding a predicted PET set T_y and non-PET set T_n.
07: Annotate T_y, yielding T'_y
08: Select m tweets randomly from predicted non-PET set T_n and annotate them, yielding T'_n
09: Add T'_y and T'_n to the training set *T*, yielding a new training set: $T \leftarrow T + T'_y + T'_n$
10: Train classifier(s) with *T*
11: **Loop**

12: **Return** T

where l is greater than m. β is the balance ratio, the ratio between the number of PET and non-PET tweets, δ is the expected accuracy. The value of m is only a fraction of the number of tweets in the newly predicted PET class (Step 06). Both l and m can be constants. The accuracy of a classifier is measured by the ROC Area and /or F-measure.

3.2 Dataset

Using the above algorithm, we constructed a PET corpus related to 4 dietary supplements: Echinacea, Melatonin, St. John's Wort, and Valerian. A total of 108,528 tweets were collected from May 30, 2014 to December 8, 2014, through the use of Twitter REST API. The supplement names were used as keywords to perform Twitter searches. The breakdowns of the collected Twitter data are: 9,210 tweets for Echinacea, 81,915 for Melatonin, 3,176 for St. John's Wort, and 14,227 for Valerian. The collected Twitter data were preprocessed to remove retweets and non-English tweets.

3.3 Features

Two types of features were used by the machine learning-based filter: metadata and textual. Metadata features are features about the tweet itself but not the text. They include user id and Twitter client application. Textual features are the ones extracted directly from the 140 character Twitter text. Most of the tweets collected were unrelated to personal experience. They were usually marketing or promotion tweets or just facts of what a supplement does. According to a study of 106 million tweets with 4262 trending topics, Kwak et al. (2010) found that the majority of the messages were news specific. In another study, Krieck and colleagues found that news information normally repeats official information and has no contribution to the early detection of disease outbreaks (Krieck et al., 2011).

It has been observed that personal pronouns appear frequently in social media posts related to personal experiences (Elgersma and de Rijke, 2008; Jiang and Zheng, 2013). Personal pronouns were considered as a feature to classify personal and impersonal sentences (Li et al., 2010).

Our observation revealed that words or phrases commonly used in one class but not in the opposite class may contribute to the accurate prediction of PET and non-PET tweets. These words or phrases were found in both tweet texts and Twitter user names - unlike the Twitter screen name, a Twitter user name can be a phrase. For example, online stores may use in their names terms such as shop, store, and market. Presence of any of such words can provide classifiers a hint to identify promotional tweets.

A client application is the software application a Twitter author uses to post Twitter messages. Westman and colleagues observed that personal tweets were more often posted from the Twitter website (Westman and Freund, 2010).
The followings are the features used in this study.

1. Occurrences of automatically categorized frequent terms in username in PET class.
2. Occurrences of automatically categorized frequent username in non-PET class
3. Count of URLs in a tweet
4. Count of emotion words in a tweet
5. Count of unique words in a tweet
6. Total word count of a tweet
7. Occurrences of frequent words in PET class
8. Occurrences of frequent words in non-PET class
9. Count of pronouns in a tweet
10. Count of personal pronouns in a tweet
11. Count of first person pronouns in a tweet
12. Count of second person pronouns in a tweet
13. Count of third person pronouns in a tweet
14. Count of singular proper nouns in a tweet
15. Count of automatically categorized frequent terms in PET class
16. Count of automatically categorized frequent terms in non-PET class
17. Occurrences of frequent terms in Twitter user name
18. Client application used to post the tweet
19. Twitter user id

3.4 Classifiers

For filtering the off-topic tweets, three classifiers were used: decision tree (J48), KNN (IB1) and, neural network (Multilayer Perceptron, MLP). Neural networks are known for deriving meaning from complex and imprecise data. Decision trees are simple to understand, interpret and, easily handle feature interaction. KNN is simple and robust for noisy data. For evaluation purposes, both ROC metrics and F-measure were used for the reason that F-measure is not an appropriate measure of performance when the data are imbal-

anced (Chawla, 2009). Weka (Hall et al., 2009) which contains the implementation of all three algorithms was used in our study. It is well understood that not all classifiers perform the same way. The majority rule was used to determine the outcome of classification. That is, if outputs of two or more classifiers were PET, then the tweet was considered a PET tweet.

4 Results

Using a seed of 3,176 tweets (Run 0), our algorithm had gone three iterations with the test sets shown below. In each iteration (Run 1 through Run 3), the size of training set (corpus) increased as more annotated tweets were added.

Iteration	Training Set	Test Set	# Predicted PET Tweets	# Non-PET Tweets Added
Run 1	3,176	9,210	94	31
Run 2	3,301	14,277	386	128
Run 3	3,815	81,915	3,721	1,235

Table 1: Dataset size over iterations. It shows the number of tweets in the training set, test set, predicted PET set, and added non-PET set in each iteration.

The final annotated data set consisted of 8,770 number of tweets which are available at `https://github.com/medeffects/supplement-corpus/` . Of these, 2,067 were PET tweets and 6,703 non-PET tweets.

4.1 Inter-Annotator Agreement

Inter-annotator agreement metrics are helpful to establish the subjectivity of an annotation scheme. The annotation task was performed by 2 annotators. Two labels were used for the annotation: PET and non-PET. As shown in the table below, the average agreement was 85.4 %. Correcting for expected chance agreement, *kappa* and the other metrics still provide a reasonable score to assess the annotation consistency. The result indicates that the task of finding personal experience tweets does have a level of subjectivity. These values can later be useful to define an expected upper boundary on the PET classification task.

4.2 Corpus Class Balance

As stated earlier, the corpus was built in iterations (or runs). Each iteration used a larger training set that consisted of more examples of PET tweets.

kappa	0.624
alpha	0.624
Average Agreement	0.854
π	0.624
S	0.806

Table 2: Inter-annotator agreement for metrics

As such, it can be noticed that with each iteration more PET tweets were added to the corpus as shown in Table 3, leading to a more balanced distribution of PET and non-PET tweets. This result is beneficial for this study since the goal of it is to find as many personal experience tweets which can later be used to associate effects with dietary supplements for health surveillance.

Iteration	PET	Non-PET	Ratio
Run 0	98	3,078	1:31
Run 1	145	3,156	1:22
Run 2	256	3,559	1:14
Run 3	2,067	6,703	1:3

Table 3: Corpus class balance over iterations

4.3 Classifier Performance

In addition to studying the overall performance of classifiers collectively, we also collected performance data of each individual classifier on predicting PET tweets, and they are shown in the figure below.

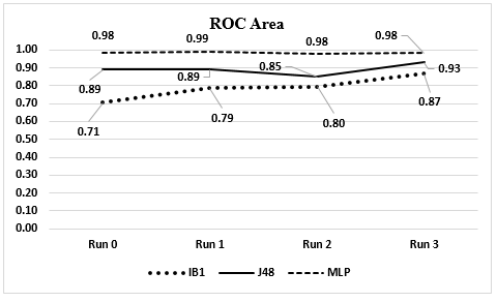

Figure 1: ROC area over iterations

4.4 Feature Ranking

One important aspect in this study is to determine what features helped to automatically detect personal experience tweets. As indicated previously, most of these 19 features by classifiers were extracted from the tweet text using natural language processing techniques. To perform the feature analysis, the Chi-Square ranking method was used. The top ranked features are occurrences of

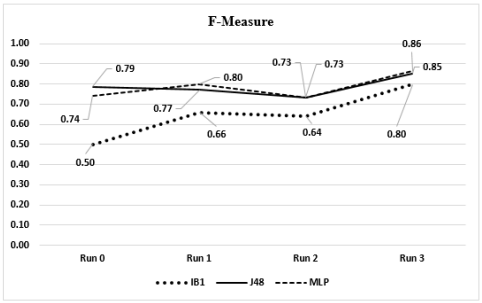

Figure 2: F-Measure over iterations

automatically categorized frequent terms in username in PET class, occurrences of automatically categorized frequent username in non-PET class, occurrences of frequent words in PET class, occurrences of frequent words in non-PET class, pronoun count, personal pronoun count, first person pronoun count, URL count, Twitter client and user id.

4.5 Prediction Precision

The overall performance of the PET classifiers was measured with the training sets. The PET classifiers are the filter used to identify relevant tweets for human annotation. The actual performance of the filter should be measured against the prediction using the test data. Given that only predicted PET tweets were annotated –that is, only true positive and false positive figures were available, prediction precision was measured. Precision is a ratio between actual PET tweets and predicted PET tweets in the same predicted PET set, a performance measurement of classifiers when performing predictions.

Table 4 shows that the precision falls within the range of 0.28 - 0.49. This indicates that for every 100 predicted samples (tweets), between 28 and 49 may be actual PETs.

# PET Tweets			
Iteration	Predicted	Actual	Precision
Run 1	94	46	0.49
Run 2	386	107	0.28
Run 3	3,721	1,597	0.43

Table 4: Prediction precision over iterations

5 Discussions

The amount of work on annotation can be significant when constructing a corpus that requires examination of large sets of data. In this study, if we were to annotate 108,528 tweets, it would take annotators a significant amount of their time to do so. However, using our proposed method, two annotators only needed to annotate 8,770 tweets (= initial seed tweets plus predicted PET tweets and added non-PET tweets in each iteration. Refer to Table 1). If it takes an average of one minute to annotate a single tweet and each annotator spends 8 hours a day on annotation, it will take an annotator 226 days to complete annotation of 108,528 tweets, but 18 days for 8,770 tweets. This represents a significant reduction of annotation time.

By some estimates, the obtained kappa score shown in Table 2 may be considered low which implies that the text is highly subjective and difficult to annotate. This suggests that finding personal experience tweets is highly subjective. Personal experience, in the context of this paper, is text expressed by a person and that is of a very personal nature. The difficulty may lie in the fact that there is not set lexicon to define personal experience. In contrast, emotion text detection, which is also considered subjective, does have its own lexicon (i.e happy words vs. sad words).

As can be seen in Table 3, our approach is also efficient in improving the class balance of the corpus. With only 3 iterations, the ratio of the number of PET tweets to that of non-PET tweets had come down from 1:31 to 1:3, a 10-fold improvement.

The performance of individual classifiers on predicting PET tweets with the training data either remained the same level or improved over iterations. For ROC Area (Figure 1), both IB1 and J48 improved, and MLP remained the same. For F-Measure (Figure 2) which is not an appropriate indicator of performance when data are imbalanced, all three classifiers had improved. In addition, it is noted that the multilayer perceptron (MLP) classifier has the best accuracy in predicting PETs.

Although values of ROC Area and F-Measure are quite promising, when it came to predict the unlabeled data (test set), 3 classifiers could only predict PET tweets with 28% to 49% precision. This implies that if the classifiers are to be used to predict PETs on new sets of unlabeled tweets, only 28% to 49% of tweets in the predicted PET set may be actual PET tweets.

Our result of feature ranking suggests that between metadata and textual features, textual features contribute the most to overall classification

accuracy. And the best performing features are the ones related to the frequency of terms used in either tweet text or the user name - that is, the most frequent terms in a class that are infrequent in the opposite class. This approach is sometimes commonly referred to as the Gramulator type approach.

6 Conclusion

We proposed a bootstrap method to construct tweet corpus from noisy Twitter data. Through a few iterations, our approach can help construct quickly a tweet corpus with closely balanced classes, without a significant amount effort on annotation. It is conceivable that our approach can be applied to other health surveillance studies that use machine learning-based classifications of imbalanced social media data.

Acknowledgment

Authors wish to thank anonymous reviewers for their significant effort in critiquing our work and providing constructive comments, Yongbing Tang for data collection for this project, Jiabao Liu for coding for data processing and analysis, and Cecelia Lai for annotating the tweets. This work was supported in part by the National Institutes of Health grant 1R15LM011999-01.

References

Jiang Bian, Umit Topaloglu, and Fan Yu. 2012. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM.

Ugur Bilge, Selen Bozkurt, Basak Oguz Yolcular, and Deniz Ozel. 2012. Can social web helpoff detect influenza related illnesses in turkey? *Stud Health Technol Inform*, 174:100–104.

David A Broniatowski, Michael J Paul, and Mark Dredze. 2013. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672.

Michael Chary, Nicholas Genes, Andrew McKenzie, and Alex F Manini. 2013. Leveraging social networks for toxicovigilance. *Journal of Medical Toxicology*, 9(2):184–191.

Nitesh V Chawla. 2009. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer.

Cynthia Chew and Gunther Eysenbach. 2010. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118.

Rumi Chunara, Jason R Andrews, and John S Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American journal of tropical medicine and hygiene*, 86(1):39–45.

Nigel Collier, Nguyen Truong Son, and Ngoc Mai Nguyen. 2011. Omg u got flu? analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, 2(5):1.

Preciosa M Coloma, Benedikt Becker, Miriam CJM Sturkenboom, Erik M van Mulligen, and Jan A Kors. 2015. Evaluating social media networks in medicines safety surveillance: two case studies. *Drug safety*, 38(10):921–930.

Mark Dredze. 2012. How social media will change public health. *Intelligent Systems, IEEE*, 27(4):81–84.

Erik Elgersma and Maarten de Rijke. 2008. Personal vs non-personal blogs: initial classification experiments. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 723–724. ACM.

Clark C Freifeld, John S Brownstein, Christopher M Menone, Wenjie Bao, Ross Filice, Taha Kass-Hout, and Nabarun Dasgupta. 2014. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug safety*, 37(5):343–350.

Isaac Chun-Hai Fung, King-Wa Fu, Yuchen Ying, Braydon Schaible, Yi Hao, Chung-Hong Chan, and Zion Tsz-Ho Tse. 2013. Chinese social media reaction to the mers-cov and avian influenza a (h7n9) outbreaks. *Infectious diseases of poverty*, 2(1):1–12.

Francesco Gesualdo, Giovanni Stilo, Michaela V Gonfiantini, Elisabetta Pandolfi, Paola Velardi, Alberto E Tozzi, et al. 2013. Influenza-like illness surveillance on twitter through automated learning of naïve language. *PLoS One*, 8(12):e82489.

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen OConnor, Abeed Sarker, Karen Smith, and Graciela Gonzalez. 2014. Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark. In *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*. Citeseer.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an

update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. 2013. Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of medical Internet research*, 15(4):e62.

Jenine K Harris, Raed Mansour, Bechara Choucair, Joe Olson, Cory Nissen, Jay Bhatt, et al. 2014. Health department use of social media to identify foodborne illness-chicago, illinois, 2013-2014. *MMWR Morb Mortal Wkly Rep*, 63(32):681–685.

N Heaivilin, B Gerbert, JE Page, and JL Gibbs. 2011. Public health surveillance of dental pain via twitter. *Journal of dental research*, 90(9):1047–1051.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis*.

Keyuan Jiang and Yujing Zheng. 2013. Mining twitter data for potential drug effects. In *Advanced data mining and applications*, pages 434–443. Springer.

Manuela Krieck, Johannes Dreesman, Lubomir Otrusina, and Kerstin Denecke. 2011. A new age of public health: Identifying disease outbreaks by analyzing tweets. In *Proceedings of Health Web-Science Workshop, ACM Web Science Conference*.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Shoushan Li, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. 2010. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 414–423. Association for Computational Linguistics.

Kimberly McManus, Emily K Mallory, Rachel L Goldfeder, Winston A Haynes, and Jonathan D Tatum. 2015. Mining twitter data to improve detection of schizophrenia. *AMIA Summits on Translational Science Proceedings*, 2015:122.

Ruchit Nagar, Qingyu Yuan, Clark C Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and John S Brownstein. 2014. A case study of the new york city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. *Journal of medical Internet research*, 16(10):e236.

Anna C Nagel, Ming-Hsiang Tsou, Brian H Spitzberg, Li An, J Mark Gawron, Dipak K Gupta, Jiue-An Yang, Su Han, K Michael Peddecord, Suzanne Lindsay, et al. 2013. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *Journal of medical Internet research*, 15(10):e237.

Thiago D Nascimento, Marcos F DosSantos, Theodora Danciu, Misty DeBoer, Hendrik van Holsbeeck, Sarah R Lucas, Christine Aiello, Leen Khatib, MaryCatherine A Bender, Jon-Kar Zubieta, et al. 2014. Real-time sharing and expression of migraine headache suffering on twitter: A cross-sectional infodemiology study. *Journal of medical Internet research*, 16(4):e96.

Michelle Odlum and Sunmoo Yoon. 2015. What can we learn about the ebola outbreak from tweets? *American journal of infection control*, 43(6):563–571.

Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from twitter. *Health*, 11:16–6.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467.

Mustafa Sofean and Matthew Smith. 2012. Sentiment analysis on smoking in social networks. *Studies in health technology and informatics*, 192:1118–1118.

Stina Westman and Luanne Freund. 2010. Information interaction in 140 characters or less: genres on twitter. In *Proceedings of the third symposium on Information interaction in context*, pages 323–328. ACM.

Elad Yom-Tov, Diana Borsa, Ingemar J Cox, and Rachel A McKendry. 2014. Detecting disease outbreaks in mass gatherings using internet data. *Journal of medical Internet research*, 16(6):e154.