

Did You Really Just Have a Heart Attack? Towards Robust Detection of Personal Health Mentions in Social Media

Payam Karisani
Emory University
payam.karisani@emory.edu

Eugene Agichtein
Emory University
eugene.agichtein@emory.edu

ABSTRACT

Millions of users share their experiences on social media sites, such as Twitter, which in turn generate valuable data for public health monitoring, digital epidemiology, and other analyses of population health at global scale. The first, critical, task for these applications is classifying whether a personal health event was mentioned, which we call the (*PHM*) problem. This task is challenging for many reasons, including typically short length of social media posts, inventive spelling and lexicons, and figurative language, including hyperbole using diseases like “heart attack” or “cancer” for emphasis, and not as a health self-report. This problem is even more challenging for rarely reported, or frequent but ambiguously expressed conditions, such as “stroke”. To address this problem, we propose a general, robust method for detecting PHMs in social media, which we call *WESPAD*, that combines lexical, syntactic, word embedding-based, and context-based features. *WESPAD* is able to generalize from few examples by automatically distorting the word embedding space to most effectively detect the true health mentions. Unlike previously proposed state-of-the-art supervised and deep-learning techniques, *WESPAD* requires relatively little training data, which makes it possible to adapt, with minimal effort, to each new disease and condition. We evaluate *WESPAD* on both an established publicly available Flu detection benchmark, and on a new dataset that we have constructed with mentions of multiple health conditions. Our experiments show that *WESPAD* outperforms the baselines and state-of-the-art methods, especially in cases when the number and proportion of true health mentions in the training data is small.

KEYWORDS

Social media classification; Health tracking in social media; Representation learning for text classification.

ACM Reference Format:

Payam Karisani and Eugene Agichtein. 2018. Did You Really Just Have a Heart Attack? Towards Robust Detection of Personal Health Mentions in Social Media. In *Proceedings of The 2018 Web Conference (WWW 2018)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3178876.3186055>

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW 2018, April 23-27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5639-8/18/04.

<https://doi.org/10.1145/3178876.3186055>

1 INTRODUCTION

Individuals and organizations increasingly rely on social data, created on platforms such as Twitter or Facebook, for sharing information or communicating with others. Large volumes of this data have been available for research, opening new opportunities to answer questions about society, language, human behavior, and health. Among these, monitoring and analyzing social data for public health has been an active area of research, due to both the importance of the topic, and to the unprecedented opportunities afforded by a real-time window into the self-reported experience of millions of people online. These social data come with many challenges and potential biases[31]. Nevertheless, these data already enabled many public health applications, such as tracking the spread of influenza[2, 32], understanding suicide ideation[10], monitoring and providing support during humanitarian crises[17], drug use [13, 35], drinking problems [24], and public reactions to vaccination [37].

The main advantages of social data over traditional methods of public health surveillance such as phone surveys, in-person interviews, and clinical reports, include scalability and potential near-real time responsiveness. Therefore, social data has become a valuable source to monitor and analyze people’s reports and reactions to health-related incidents. A crucial first step in disease analysis and surveillance using social data, is to identify whether a user post is actually mentioning a specific person reporting a health event. All subsequent processing and analysis, whether it is epidemic detection (e.g., mentioning an affected person in the post), or individual analysis (e.g., reporting one’s own health condition), depends on the accuracy of the detection and categorization of the individual postings. If the posting were mis-categorized, and did not in fact report a health-related event, all subsequent analysis and conclusions arising from the data might be flawed.

Our goal is to accurately identify postings in social data, which not only contain a specific disease or condition, but also mention a person who is affected. For instance, we aim to identify posts such as: “My grandpa has Alzheimer’s & he keeps singing songs about wanting to forget” or “Yo Anti-Smoking group that advertises on twitch, I don’t smoke. My mom died to lung cancer thanks to smoking for like 40 years. I get it.”. In contrast, we wish to filter out postings like: “I almost had a heart attack when I found out they’re doing a lettering workshop at @heathceramics in SF” or “Dani seems like a cancer, spreads herself anywhere for attention!”. In terms of previous work, we aim to identify specific health reports, rather than non-relevant postings or postings expressing general concern or awareness of a disease or condition [22]. We call this task detecting Personal Health Mentions, or *PHM*. Further, we aim to develop a solution that is both robust and general, so that it can scale to many diseases or conditions of current or future interest. In turn, more accurately

detecting personal health mentions in social data, without requiring extensive disease-specific development and tuning, would empower public health researchers, digital epidemiologists and computational social scientists to ask new questions, and to answer them with higher confidence.

Detecting health mentions in social data is a challenging task. Social data posts, such as those on Twitter, tend to be short, and are often written informally, using diverse dialects, and inventive and specialized lexicons. Previous efforts for similar tasks applied machine learning methods that relied on extensive feature engineering, or on external feature augmentation to address the sparsity in the feature space, e.g., for company name detection [40], reputation measurement [5], sarcasm detection [3, 18], and for public health [9, 22]. In the health context, the problem is exacerbated by the limited availability of training data, and by the low frequency of the health reports in even keyword-based samples: on Twitter, our experiments show that of the tweets containing a disease name keyword, only 19% are actual health reports. The resulting classifiers tend to have high precision, but relatively low recall (i.e., high false negative rate), which may not be desirable for applications such as disease surveillance or detecting epidemics.

Our goal is to address the problems of sparsity and imbalanced training data for *PHM* detection, by explicitly modeling the differences in the distribution of the training examples in the word embeddings space. For this, we introduce a novel social text data classification method, *WESPAD* (**W**ord **E**mbedding **S**pace **P**artitioning and **D**istortion), which learns to *partition* the word embeddings space to more effectively generalize from few training examples, and to *distort* the embeddings space to more effectively separate examples of true health mentions from the rest. While deep neural networks have been proposed for this purpose, our method works well even with small amounts of training data, and is more simple and intuitive to tune for each new task, as we show empirically in this paper. We emphasize that *WESPAD* requires no topic or disease-specific feature engineering, and fewer training examples than previously reported methods, while more accurately detecting true health mentions. These properties make *WESPAD* particularly valuable for extending public health monitoring to a wider range of diseases and conditions. Specifically, our contributions are:

- We propose a novel, general approach to discover domain-specific signals in word embeddings to overcome the challenges inherent in the *PHM* problem.
- We implement our approach as an effective *PHM* classification method, which outperforms the state-of-the-art classifiers in the majority of settings.
- To validate our approach, we have constructed and released a manually-annotated dataset of Twitter posts for six prominent diseases and conditions¹.

Next, we review related work to place our contributions in context.

2 RELATED WORK

Social network data and user-contributed posts on platforms such as Facebook and Twitter, have been extensively studied for diverse applications in business, politics, science, and public health. Some

prominent examples include work on answering social science questions [23], and analyzing influenza epidemics [7]. Our work builds on three general directions in this area: general classification techniques that serve as the foundation of our work; disease-specific classifiers for social text data; and, closest to our work, prior research on general health classifiers that could be potentially applied to different diseases and conditions.

2.1 Text Classification: Models and Techniques

Methods for automatic text classification have been studied for decades, and have evolved from simple bag-of-words models to sophisticated algorithms incorporating lexical, syntactic, and semantic information [1]. Many of these algorithms have been adapted for biomedical text processing, with varying success [11, 33]. Recently, deep neural networks have emerged in many areas of natural language processing as an alternative to feature engineering and demonstrating new state-of-the-art performance on a wide range of tasks. However, there are two main challenges in making these models effective. First, it is commonly known that deep neural models usually need a large amount of training data to reach their ultimate capacity. This is an active area of research, and workshops such as Limited Labeled Data (LLD)² are held to investigate this area. Second, it is not clear how to incorporate domain knowledge into the training—e.g., social network topology, or user activities. Nevertheless, we include three state-of-the-art deep neural network models as baselines, chosen as representative of the most effective neural network methods reported for text classification. We show empirically that our proposed method performs better than these techniques, especially in the settings with small amounts of available training data.

To improve the generalization of classification to unseen textual cases, *word embeddings* [28] have been proposed as a semantic, and broader, representation of text. This idea has been used, for example, to compare two sentences, in addition to lexical features. For instance, [4] proposed a system for detecting semantic similarity between two pieces of texts using word embeddings similarity. [19] proposed an algorithm similar to [4], for paraphrase detection task. Their main contribution is that the algorithm can capture the similarity between two sentences with higher details through binning the word similarity values. For another task, sarcasm detection, reference [18] evaluated a number of word embeddings features to discover word incongruity, by measuring the similarities between the word vectors in the sentence. While these studies have been done for different domains and tasks, they all share the same idea of using word embeddings space as a resource to extract features; we also build on this general idea for the *PHM* detection problem. Reference [44] proposed the idea of clusters of word vectors to address the problems of word ambiguity. The clusters are used to generate compound embeddings features. In our experiments we observed that word clusters do not accurately characterize social data texts when used directly. Instead, we propose *partitioning* the word embeddings space to generate features describing the distribution of training examples in the different regions of the space.

¹The dataset and code are available at <https://github.com/emory-irlab/PHM2017>

²Available at <https://lld-workshop.github.io/>

The resulting distributions are subsequently used to map the instances of each class to different categories, which, as we show allows *WESPAD* to more precisely identify true instances of *PHM*.

2.2 Text Classification for Health

Disease-specific text classifiers: Since building a large training set for public health monitoring is costly, and in some cases impossible (e.g., for rarely reported diseases); it has been shown that domain knowledge in the form of rule-based, or domain-specific classifiers is effective in monitoring certain diseases, e.g., [9, 22]. A large body of work has been done for detecting and tracking information about specific diseases. This includes investigations on tracking the spread of flu epidemics [2, 22], cancer analysis [30], asthma prediction [12], depression prediction [9, 42], and anorexia characterization [8]. To improve accuracy for each of these domains, studies such as [22] have shown that certain aspects of tweets are also good indicators of health reports, and have been successfully operationalized as lexical and syntactic features, which we incorporate into our baseline system. A thorough overview of the published papers can be found in references [6, 33]. Our work can be potentially used to improve the accuracy of health mention detection for *all* of the mentioned disease specific studies. We emphasize that an advantage of our model, introduced in the next section, is that without imposing any restriction on the original features, our method can substantially improve the detection accuracy, even when there is only a small set of positive examples available.

General-purpose text classification for Health A more attractive strategy than developing disease-specific classifiers, is to develop a single classification algorithm, that could be easily adapted to detect mentions of different diseases and conditions. This is the direction we chose in this work. Reference [32] reports using an LDA topic-based model, which also incorporates domain knowledge, to discover the symptoms and their associated ailments in Twitter. [36] proposed a two step process, which is representative of a common methodology, to detect the health mentions in social text data. The first, high-recall step is to collect the tweets using keywords and regular expressions, and the second step is to use a high-precision classifier – in this case, by using a correlation-based feature extraction method. Reference [43] reported using a dataset of tweets across 34 health topics, and investigated the accuracy of the classifiers trained over multiple diseases and tested on new diseases. The authors conclude that training a classifier on four diseases: cancer, depression, hypertension, and leukemia can lead to a general health classifier with 77% accuracy using standard SVM classifiers and bag-of-words features, similar to one of our baselines used in empirical evaluation, which, as we will show, is not able to generalize well to unseen test data. We emphasize that our aim is also to develop a general health mention detection model that could apply to a variety of diseases and conditions.

In summary, to our knowledge, our *WESPAD* model (presented next) is the first *general* health report detection method that requires only small amounts of training data, does not do any domain-specific feature engineering, yet performs as well as, and often better than, other methods, including a disease-specific rule-based classifier.

3 WESPAD MODEL DESCRIPTION

This section introduces our method, *WESPAD*, for robust classification of health mentions in social data. We first summarize previously proposed lexical and syntactic features for social media classification, used both for health mentions, and other domains, which we use as starting point for our method. We then introduce the novel steps of our work, for learning topic-specific representation of the data derived from word embeddings (Sections 3.3 and 3.4).

3.1 Lexical and Syntactical Features

Previous studies on analyzing social media (primarily Twitter and Facebook posts) for depression prediction [9], influenza tracking [22], and tobacco use [35], have shown that certain words and phrases are key indicators of the health reports. Therefore, we use all word unigrams and bigrams as features, in order to capture any words or phrases that may be salient.

Additionally, to model syntactic dependencies in the text, we use the approach proposed in [26] to identify common syntactical dependencies in the tweets through detecting the frequent syntactical subtrees, and use them as features. To detect the sentences in the posts, we used the tweet dependency trees [21] to detect sentence boundaries. We conjecture that even with small amount of training data, the frequent subtrees can automatically detect a subset of syntactic patterns which are usually designed manually for certain health cases (such as those of flu detection in [22]). Our experiments (in Section 6) show that lexical and syntactic features provide high precision for health mention detection, but are not sufficient to generalize from the (relatively small) amounts of training data. To improve generalization, we now describe our use of word embeddings, which allows learning from a few positive examples of health mentions. In the next sections, we use word *lex_feats* to refer to the bigrams, and use word *syn_feats* to refer to the features extracted from the frequent subtrees.

3.2 Detecting “Noisy” Regions in the Word Embeddings Space

Word embeddings [28] is an approach to map words to linguistic concepts in a lower-dimensional vector space. The motivation for using word embeddings to address sparsity in our task is that it could help match and generalize training examples to unseen examples at test time, which may not share the same words but have semantically related meaning. A common way to represent a short piece of text in the word embeddings space is to average of the constituent word vectors, and use the centroid of the vectors directly as features for a classifier. Although this approach has some drawbacks, e.g., losing information about individual words, several studies have shown that it can be effective [4, 19, 39]. Here we explore the ways that the centroid representation can be extended to improve classifier generalization to unseen cases.

One could incorporate the classifier output (predicted class) alongside other features in the final feature vector. However, as we will show, incorporating the classifier output, as is, would propagate the false positive matches in the word embeddings space to generate noisy features for the final classifier. Another problem with the approach above is that a centroid in the word embeddings

space does not preserve information about the constituent words, and thus likely to map both positive and negative examples of a health mention (if they share common words) into similar vectors in the word embeddings space. However, we will use the centroids as a starting point, in combination with a classifier trained over the centroid features, to detect, and downweight, the regions in the word embeddings space, where the positive and negative training examples have such a similar centroid vectors that they are no longer distinguishable.

Definition: “Noisy” regions in the word embeddings space: We define “noisy” regions as those, where the precision of a centroid-based classifier is lower than a certain threshold α .

Using this definition, we can now *filter* out the noisy regions (and the corresponding features from training data). Detecting the regions which are *not* noisy also can help us to address the challenge of an imbalanced training set. Examples with centroids mapped to these regions can be directly used in the model to predict the label of the instances which are not present in the training set. This can lead to a model which can generalize better with only a small set of positive cases. To detect the noisy regions in the embeddings space we define a probabilistic function Pr to be the probability of assigning the tweet centroid to the positive class. The values associated with Pr can be extracted from the training set using the logistic regression model as the centroid-based classifier. Given the function Pr and the associated values, the probability of assigning tweet t_i to the positive class would be $Pr(t_i)$. Based on the definition, in the noisy regions the value of Pr is close to 0.5. More formally, we define binary features $PFlag(t_i)$ and $NFlag(t_i)$ for tweet t_i as follows:

$$PFlag(t_i) = \begin{cases} 1 & 0.5 + \alpha \leq Pr(t_i) \\ 0 & \text{Otherwise} \end{cases}$$

$$NFlag(t_i) = \begin{cases} 1 & 0.5 - \alpha \geq Pr(t_i) \\ 0 & \text{Otherwise} \end{cases}$$

in which α is the threshold to detect the noisy regions, and can be tuned in the training phase. If tweet t_i is predicted to be positive and is not located in a noisy region, the value of $PFlag(t_i)$ is set, and likewise, if it is predicted to be negative, and is not located in the noisy regions the value of $NFlag(t_i)$ is set. The output of one example of the noisy region detection is illustrated in Figure 1(a). Figure 1(a) illustrates a 2-dimensional projection of positive and negative examples (marked with ‘x’ and ‘o’, respectively) using t-SNE [25], and the corresponding noisy region, the circle area where the centroids of positive and negative examples are not distinguishable with high confidence. All the data points contain word *heart attack*.

3.3 Partitioning the Word Embeddings Space

The “noisy region” flags $PFlag$ and $NFlag$ can help us capture the semantic similarity between the tweets which potentially belong to the same class. However, even though we can control the degree of uncertainty in $PFlag$ and $NFlag$ through the parameter α , they may still propagate the noise in the embeddings space to the final feature vectors. Since the original lexical feature vectors are sparse, $PFlag$ and $NFlag$ may be awarded a high weight by the final classifier,

and potentially cause more errors. To reduce the effect of these features, and also to utilize the association between the lexical features and their representation in the embeddings space, we constrain $PFlag(t_i)$ and $NFlag(t_i)$ to the region in the embeddings space in which t_i is located. Thus, we expect the two features also reflect the lexical similarity to some extent (in addition to storing information about the class label). The idea is illustrated in Figure 1(b).

Figure 1(b) shows the same space that we discussed earlier with 3 hypothetical partitions, and two features for the examples mapped to each partition. Given a tweet t_i appearing in partition P_k feature $PFlag_k(t_i)$ or $NFlag_k(t_i)$ can be set. For instance, for the positive set of tweets which appear in partition P_2 , only the value of $PFlag_2$ is set, and for the negative set of tweets which appear in partition P_3 , only the value of $NFlag_3$ is set. We emphasize that this is different from the idea of clustering the embeddings space. The partitions in our case are used to represent the original posting text along with the class labels, since the tweets which are close in the embeddings space are likely to also share lexical content. On the other hand, we don’t expect to have pure partitions due to the expected overlap in the vocabulary between the negative and positive classes.

The number of partitions (K) can be tuned experimentally in the training phase. In general, we expect large partitions to improve recall, but to decrease precision. This is because larger partitions could result in a higher number of tweets to be mapped to the same pair of $PFlag$ and $NFlag$, which can potentially increase the number of detected positive cases, and also increase the chance of mislabeling the tweets. In the rest of the paper, we use the word *we_partitioning* to refer to the features proposed in this section.

3.4 Distorting the Word Embeddings Space

In Section 3.2 we tried to partially address one of the drawbacks of directly using the word embeddings centroids, which is the loss of information about the constituent words. However, this fix does not resolve the inherent problem of using the centroids in the original word embeddings space. One approach to incorporate the information about individual terms is to integrate word importance into the computation of the tweet centroid vector. For instance, to reduce the effect of the less informative words on the centroid values, [16] suggests using IDF-weighting to compute the weighted average of the word vectors in the sentence representation context. In classification context, we propose to use information gain [29] weighting to compute the centroid vector to boost the impact of the words which are effective in the classification, effectively “distorting” the word embeddings space. More formally, we compute the new, “distorted” centroid of tweet t as:

$$\vec{M}_t = \frac{\sum_{i=0}^n IG_i \times \vec{W}_i}{\sum_{i=0}^n IG_i}.$$

where \vec{M}_t is the weighted mean vector for tweet t , \vec{W}_i is the vector representation of word W_i in tweet t , and n is the length of the tweet. IG_i is the information gain of word W_i in the training set, and is computed as:

$$IG_i = \text{Entr}(D) - \left(\frac{|D_{w_i}|}{|D|} \times \text{Entr}(D_{w_i}) + \frac{|D_{\overline{w_i}}|}{|D|} \times \text{Entr}(D_{\overline{w_i}}) \right)$$

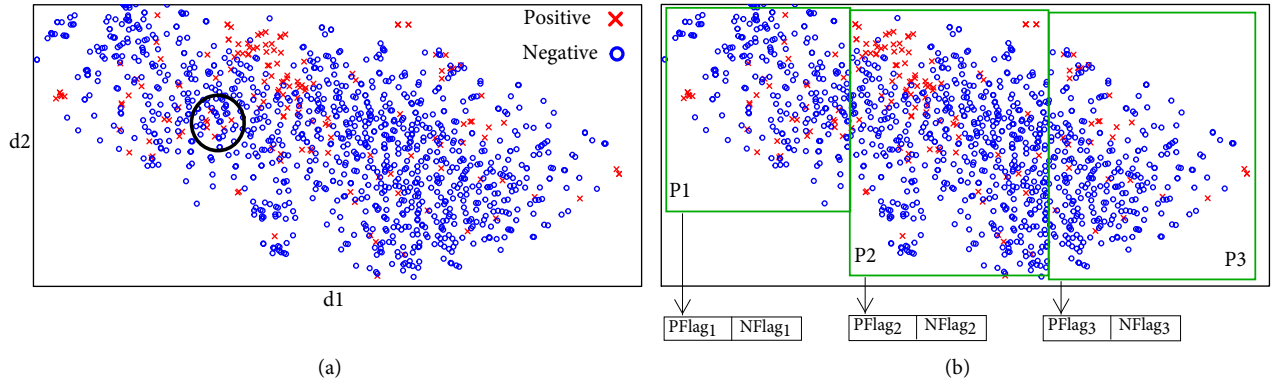


Figure 1: (a) Tweet centroids in the word embeddings space that contain the phrase *heart attack*, projected to two dimensional space using t-SNE. (b) The same word embeddings space with 3 hypothetical partitions, and a pair of features associated with each partition.

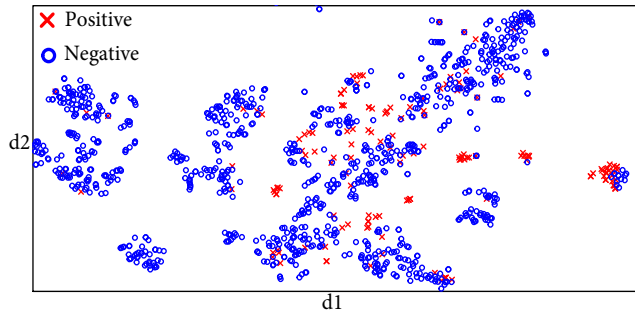


Figure 2: The same set of tweet centroids reported in Figure 1, after applying IG-weighting transformation.

where D is the training set, $Entr(D)$ is the entropy of D relative to our classification problem, $|D|$ is the size of the training set, D_{w_i} is the subset of the training set for which w_i occurs, and \bar{D}_{w_i} is the subset of the training set for which w_i does not occur. For the words which do not appear in the training set, we estimate their information gain using the information gain of their closest word in the embeddings space, that do appear in the training set.

Figure 2 shows the same set of tweets from Figure 1, after applying “IG-weighting”. The projection illustrates that in some cases, the transformation can successfully separate the tweets in different classes by mapping them to different regions of the word embeddings space.

To convert the new, weighed, centroids into features, we transform all the centroids using the information gain values extracted in the training set, and follow the model described in the previous sections to extract the centroid-based features for each tweet. The values of the parameter α and the number of partitions K , can be potentially different in the distorted word embeddings space, therefore, we call them α_2 and K_2 . In the rest of the paper, we use the term *we_distortion* to refer to the features introduced in this section.

3.5 Representing the Posting Context

Previous studies in monitoring public health have shown that user posting history is a good indicator of his or her current state, e.g., for

depression detection [9]. We hypothesize that the users who post a message which includes a true personal health mention, might have already posted or will post a similar message. Although those messages may not necessarily contain the disease keywords, they may be semantically or lexically related to the current one. Therefore, we assume that true health-related postings will be somewhat consistent with the other, contemporaneous, posts by the user. Of course, the actual effect of the health event on the user depends on a variety of factors, e.g., the severity of the condition. To enable our model to capture these effects, in a way appropriate for each disease or condition of interest, we include a representation of the prior and subsequent posts by the user³. Therefore, we use the representation described in Sections 3.2 and 3.3 to also represent the prior- and next- tweets of the user, and incorporate the resulting features into the final combined feature vector. In the subsequent sections, we use the terms *context_prev* and *context_next* to refer to the features extracted from the previous and next user messages respectively.

4 WESPAD CLASSIFIER IMPLEMENTATION

So far, the proposed representation model provides a general approach for feature learning, and can be implemented with a number of different algorithms. We now describe the specific implementation to operationalize *WESPAD* into a classifier used for experiments in the rest of the paper. We emphasize that the implementation described below is just one (effective) way to operationalize the proposed model.

Lexical and syntactic features (Section 3.1): to parse and build the dependency tree for the tweet contents we used the parser introduced in [21]. No stemming or stopword removal was performed⁴. To extract the frequent subtrees, we used the approach proposed in [26], with minimum support 10 and minimum tree size 2, as suggested in [26].

Word embeddings implementation: We experimented with multiple pre-trained word embeddings implementations. Specifically,

³Given the regular limitations in using social network API to retrieve the user postings, accessing the previous and next messages of the user might be problematic, specifically in the real-time large-scale applications.

⁴In our development experiments stemming and stopword removal was not helpful.

we compared the *word2vec* word embeddings [28] (with 300 dimensions), and the pretrained *GloVe* word embeddings [34] (with 200 dimensions), specifically trained on Twitter data. We observed that the performance was impacted relatively equally by both word embeddings; for generality, we use the “standard” available *word2vec* word embeddings⁵ for all of the reported experiments. Additional incremental improvements to our method may be achieved with further training of the word embeddings on domain-specific data, as done by some of the methods that we compare to in Section 6.

Detecting noisy regions in the word embeddings space (Section 3.2): to detect the “noisy” regions in the word embeddings space, we implement the probabilistic mapping function Pr by using the Mallet implementation [27] of the multivariate logistic regression classifier with default settings.

Partitioning the word embeddings space (Section 3.3): to partition the word embeddings space into homogeneous regions we used the ELKI [38] implementation of the K -means clustering algorithm. The value of K was chosen automatically for each task as described in Section 5.3.

Combining WESPAD features for health mention prediction:

Finally, we combine the lexical, syntactic, word embedding-based, and context features described above into a joint model. For simplicity and interpretability, we used a logistic regression classifier, trained over the final feature vectors to label the tweets⁶. Other classification algorithms, such as GBDT [14], may potentially provide additional improvements by capturing non-linear relationships between the features, and may be explored in the future work.

For simplicity, the specific WESPAD implementation described above will be simply referenced as WESPAD for all of the reported experiments in the rest of the paper.

5 EXPERIMENTAL SETUP

We now describe the datasets that we used in the experiments, which include both an established benchmark dataset, and a new dataset created for the evaluation. Then we describe the baseline methods, and the experimental setup used for reporting and analyzing the results.

5.1 Datasets

We used two datasets for training and evaluation. First, the prominent benchmark dataset introduced in reference [22], focusing on identifying reports of influenza infection. This dataset, which we call *FLU2013* serves for calibration and benchmarking of our method and others, against a state-of-the-art method specifically designed for detecting Flu infection reports [22]. To explore the scalability of the PHM detection of multiple diseases and conditions, we also created a new dataset, PHM2017, described below.

FLU2013: this dataset was introduced in [22], and focused on separating *awareness* of the disease from actual *infection* reports. Each tweet in the dataset was manually labeled into classes of flu awareness (negative) or flu report (positive). Since only Twitter IDs were

⁵Available at <https://code.google.com/archive/p/word2vec/>.

⁶We also experimented with an SVM classifier with linear kernel, and initially achieved slightly better results on development data. However, the improvement came at the cost of higher training time, therefore, we opted to stay with the simpler logistic regression model.

Topic	Tweet count	self-mention	other-mention	awareness	non-health
Alzheimer	1256	1%	17%	80%	2%
heart attack	1219	4%	9%	17%	70%
Parkinson	1040	2%	9%	65%	24%
cancer	1242	3%	18%	62%	17%
depression	1213	37%	3%	49%	11%
stroke	1222	3%	11%	29%	57%

Table 1: The distribution of tweets over topics and labels in PHM2017 dataset.

distributed, the content of the tweets had to be retrieved for this study, which was done in winter 2017. At that time, there were 2,837 tweets still available to download, which is 63% of the original dataset. There were 1,393 awareness (negative class) tweets, which account for 49% of the dataset, and 1,444 report (positive class) tweets which account for 51% of the dataset.

PHM2017: We also constructed a new dataset consisting of 7,192 English tweets across six diseases and conditions: Alzheimer’s Disease, heart attack (any severity), Parkinson’s disease, cancer (any type), Depression (any severity), and Stroke. We used the Twitter search API to retrieve the data using the colloquial disease names as search keywords, with the expectation of retrieving a high-recall, low precision dataset. After removing the re-tweets and replies, the tweets were manually annotated. The labels are:

- *self-mention*. The tweet contains a health mention with a health self-report of the Twitter account owner, e.g., “*However, I worked hard and ran for Tokyo Mayer Election Campaign in January through February, 2014, without publicizing the cancer.*”
- *other-mention*. The tweet contains a health mention of a health report about someone other than the account owner, e.g., “*Designer with Parkinson’s couldn’t work then engineer invents bracelet + changes her world*”
- *awareness*. The tweet contains the disease name, but does not mention a specific person, e.g., “*A Month Before a Heart Attack, Your Body Will Warn You With These 8 Signals*”
- *non-health*. The tweet contains the disease name, but the tweet topic is not about health. “*Now I can have cancer on my wall for all to see <3*”

In our experiments, *self-mention* and *other-mention* labels are taken as positive class; and *awareness* and *non-health* labels are taken as negative class. To validate the labels, we engaged another annotator and randomly re-annotated 10% of the tweets for each topic. Since we observed that the probability of having a disputed positive label is higher than having a disputed negative label, the 10% re-labeling subset was drawn from the positive set. The re-annotation showed 85% agreement between the annotators, which is acceptable for a challenging topic like health report. Table 1 summarizes PHM2017 dataset. We observe that for each topic, a large portion of the tweets are in *non-health* category, which shows that people tend to use these words in other contexts too—which confirms the previous findings [43]. The statistics also show that, on average 19.5% of the tweets in each topic are positive, which makes the classification task more challenging. Having both a balanced

dataset (FLU2013) and an imbalanced dataset (PHM2017) helps us to evaluate our method in different settings.

To build the context features for PHM2017 dataset, we used Twitter API to download the user timelines. We were unable to build the context features for many of the tweets in the flu dataset, since in many cases either the timeline was inaccessible, or access to the user profile was restricted. Therefore, we report the results for FLU2013 without incorporating the context features, which, as we show, are helpful in PHM2017 dataset, and are expected to be available for many applications.

5.2 Methods Compared

We implemented or adapted the following methods to compare *WESPAD* to both previously used methods for health classification, and to the latest classification methods based on deep neural networks that have shown promising performance for other tasks. In Section 2, we discussed that deep neural network classifiers need a large training set to reach their best performance; however, we included these state-of-the-art baselines to compare to the models which only rely on word embeddings.

- *ME+lex*. We used a logistic regression classifier (a.k.a Maximum Entropy classifier) trained over unigrams and bigrams.
- *ME+cen*. We used a logistic regression classifier trained over the text centroid representation of the tweets in the embeddings space.
- *ME+lex+emb*. We computed the text centroid representation of each tweet in the embeddings space, and combined the resulting vector with the unigrams and bigrams of the tweet. Then a logistic regression classifier was trained over the final vectors.
- *ME+lex+cen*. We added two features *PFlag* and *NFlag* to the corresponding vector of unigrams and bigrams of each tweet. Then we used the prediction of *ME+cen* to set the values of *PFlag* to true if predicted positive, and *NFlag* to true if predicted negative. Finally, a logistic regression classifier was trained over the resulting vectors, to evaluate the contribution of our noise filtering method (Section 3.2).
- *Rules*. Experiments in [22] suggest that manually extracted templates and features are effective in detecting flu reports. We implemented the top six set of features reported in [22], and trained a logistic regression classifier over the resulting vectors. This model was used only in FLU2013 dataset.
- *CNN*. We used the convolutional neural network classifier introduced in [20]. We used the non-static variant, which can update the word vectors in the training. Using grid search, we tuned the number of convolution feature maps from values: {50, 100, 150}, and observed that the number of features highly depends on the training data, and thus was optimized automatically using grid search for each classification task. The rest of the hyperparameters were set to the suggested values⁷.
- *FastText*. We used the shallow neural network introduced in [15], known as FastText. This model represents the documents by taking the average over the individual word vectors, and can also update the vectors during the training. To tune the model we tried values: {0.05, 0.1, 0.25, 0.5} for learning rate, and values: {2, 4} for window size. We observed that the optimal value of the

learning rate was not fixed, neither in FLU2013 nor in PHM2017. The value of window size was optimal at 4 in FLU2013, but was not fixed in PHM2017. The rest of the hyperparameters were set to the suggested values [15].

- *LSTM-GRNN*. We used the model proposed in [41], which is a two-step classifier. In the first step the model uses a long short-term memory neural network (LSTM) to produce the sentence representations, and in the second step, uses a gated recurrent neural network (GRNN) to encode the sentence relations in the document. Tweet dependency trees [21] were used to detect sentence boundaries in order to produce the sentence representations. To tune the model, we used the values: {0.03, 0.3, 0.5} for learning rate, and observed that it is optimal at 0.3 in FLU2013, but is not fixed in PHM2017. The rest of the hyperparameters were set to the suggested values in the original implementation reference [41].
- *WESPAD*: our method, described in Section 3 and implemented as described in Section 4.

5.3 Training setup

To train and evaluate all of the methods in a fair and consistent way, we used the standard 10 fold Cross-Validation in FLU2013 dataset, and within each topic of PHM2017 dataset. The results reported in the next section are the averages over the test folds. To build the folds, we preserved the original distribution of the labels, and randomly assigned the tweets to each fold. Since the set of the positive tweets is small, we kept the folds fixed across all of the cross validation experiments, to ensure that all of the methods were trained and tested in identical train/validate/test folds and thus the results can be compared directly.

We used grid search to tune the model hyper-parameters by maximizing the F1-measure in the target (positive) set. To tune the number of partitions K and K_2 in the word embedding-based features of *WESPAD* (introduced in Sections 3.3 and 3.4), we experimented with the values: {3, 4, 5}, and observed that their optimal values depended on the training data, and thus were chosen automatically for each task. To tune α and α_2 (introduced in Sections 3.2 and 3.4) we tried values: {0.05, 0.15, 0.3}, and observed the best performance for the value 0.05 in the FLU2013 dataset, and for the value 0.3, for all the topics, in PHM2017 dataset⁸.

Evaluation Metrics: Since the proportion of the positive class in PHM2017 dataset was relatively low, the accuracy of all the models was high (on average 90%), due primarily to accurately predicting the negative (majority) class—which is not as practically important as the target (positive) class (the true health mentions and the target of our study). Therefore, in the next section we report the **F1-measure**, **Precision**, and **Recall** for the positive class.

6 RESULTS AND DISCUSSION

We now report the experimental results. First, we report the main results in Section 6.1, followed by the discussion and feature analysis in Section 6.2.

⁷Available at <https://github.com/harvardnlp/sent-conv-torch>

⁸For simplicity in the grid search we set $\alpha = \alpha_2$.

6.1 Main Results

Table 2 reports F1-measure of all the models (described in Section 5.2) across the topics in PHM2017 dataset. The experiments show that our model *WESPAD* outperforms all the baselines in the majority of the topics. The substantial difference in terms of F1-measure between *ME+lex* and *WESPAD* models, shows that our model has successfully managed to learn the characteristics of the small set of the positive tweets, and to generalize better. Another observation is that model *ME+lex+cen*, which uses lexical features alongside an output of a centroid-based classifier as an additional feature (see Section 5.2 for details), is performing relatively poorly. This validates our strategy described in 3.2 and 3.3, and the need to detect and filter out the noisy regions in the word embeddings space. We can also see that *CNN*, although with small amount of training data, is working surprisingly well. On the other hand, the complex *LSTM+GRNN* model is outperformed on all the topics by our *WESPAD* classifier.

Table 3 reports the average F1-measure, precision, and recall for all the models across the six topics in PHM2017 dataset. The results show that the main improvement of *WESPAD* comes from the higher recall, i.e., detecting additional true health mentions. Table 3 also shows that the highest precision is achieved by the simple *ME+lex* model, since this model only relies on the lexical features. On the other hand, *LSTM+GRNN* has the lowest precision, and this can be attributed to the complex structure of the network which expects to be fine-tuned during the training.

Model	F1	Precision	Recall
<i>ME+lex</i>	0.572	0.834	0.462
<i>ME+cen</i>	0.530	0.819	0.429
<i>ME+lex+emb</i>	0.593	0.833	0.483
<i>ME+lex+cen</i>	0.594	0.827	0.493
<i>LSTM-GRNN</i>	0.615	0.638	0.605
<i>FastText</i>	0.630	0.802	0.538
<i>CNN</i>	0.673	0.794	0.610
<i>WESPAD</i>	0.695	0.803	0.628

Table 3: Average F1-measure, precision, and recall in PHM2017 dataset.

Table 4 reports F1-measure, precision, and recall of all the baselines in comparison to *WESPAD* in FLU2013 dataset. The results show that *WESPAD* outperforms all the baselines, even though there are considerable differences between PHM2017 and FLU2013 datasets (in terms of the proportion of the positive tweets). The results also show that *WESPAD* performs slightly better than the disease-specific *Rules* classifier, implemented according to the descriptions in reference [22]. More detailed analysis revealed that the syntactic subtrees that we use in our model, to some extent, can also automatically capture the manually designed patterns reported in [22]. It is also worth mentioning that, all the improvements of *WESPAD* model over the lexical baseline *ME+lex* in both datasets are statistically significant using paired t-test at $p < 0.05$.

The comparison between the relative improvement of *WESPAD* in PHM2017 and FLU2013 datasets shows that our model performs significantly better in PHM2017 dataset. The improvement can be attributed to the inherent differences between these two datasets,

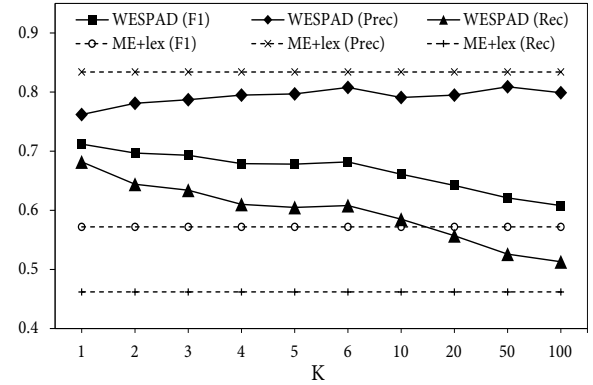


Figure 3: Impact of the number of partitions K on *WESPAD* on F1-measure, precision, and recall (PHM2017 dataset).

and the fact that PHM2017 is highly imbalanced and FLU2013 is nearly balanced. We discuss this issue further in the next section.

6.2 Discussion

We now analyze the performance of *WESPAD* in more detail, focusing on the effects of the word embeddings partitioning, contribution of different features, and the ability of *WESPAD* to generalize from few positive examples in training.

Word embeddings partitions: in Section 3.3 we argued that large partitions can increase recall, and degrade precision. To support the argument, we fixed the values of α and α_2 ; and experimented with different values for K (the number of the partitions in the regular embeddings space) and K_2 (the number of the partitions in the distorted embeddings space). Figure 3 illustrates the result of this experiment. To be able to easier interpret the results, we also set K to be equal to K_2 . The experiment confirms that by decreasing the number of partitions (and thereby increasing the partition sizes), the *Recall* of *WESPAD* improves. However, this comes at the cost of degrading the *Precision* (specifically at $K = 1$).

Feature ablation: Table 5 reports the result of the ablation study on the features in *WESPAD* model in PHM2017 dataset. The experiment shows that *we_distortion* and *we_partitioning* feature sets have the highest impact, in terms of F1-measure. We also observe that, in terms of precision, *we_partitioning* performs better than *we_distortion*. One possible explanation is that due to the small size of the positive sets, IG-weighting may fail to accurately assign the weights to the word vectors, and thus, the tweet centroid is drifted.

Effect of the number of positive examples: in Section 6.1 we observed that the relative improvement of *WESPAD* in PHM2017 dataset is considerably higher than its relative improvement in FLU2013 dataset. We argue that since FLU2013 dataset is nearly balanced, and also has a substantially larger set of positive tweets, simple models such as *ME+lex* can perform relatively well. To analyze the effect of the size of the training data, and specifically the availability of true positive examples, we varied the number of the positive examples in the training folds, by randomly sampling from 10% to 90% of the positive examples (and keeping all of the negative examples), and re-trained *WESPAD*, *Rules*, and *ME+lex* in the reduced training sets in FLU2013 dataset. Figure 4 reports the values of the F1-measure for *ME+lex*, *Rules*, and *WESPAD* at varying

Model	Alzheimer's	Heart attack	Parkinson's	Cancer	Depression	Stroke
<i>ME+lex</i>	0.701	0.399	0.468	0.533	0.722	0.610
<i>ME+cen</i>	0.704	0.327	0.383	0.587	0.727	0.453
<i>ME+lex+emb</i>	0.723	0.460	0.486	0.559	0.718	0.612
<i>ME+lex+cen</i>	0.720	0.415	0.464	0.628	0.737	0.601
<i>LSTM-GRNN</i>	0.725	0.482	0.617	0.624	0.676	0.564
<i>FastText</i>	0.769	0.491	0.540	0.605	0.741	0.633
<i>CNN</i>	0.767	0.554	0.653	0.622	0.768	0.676
<i>WESPAD</i>	0.800	0.571	0.672	0.670	0.758	0.698

Table 2: F1-measure for the models across all the topics in PHM2017 dataset.

Model	F1	Precision	Recall
<i>ME+lex</i>	0.838	0.832	0.846
<i>ME+cen</i>	0.827	0.815	0.840
<i>ME+lex+emb</i>	0.843	0.837	0.850
<i>ME+lex+cen</i>	0.844	0.843	0.845
<i>Rules</i>	0.845	0.837	0.855
<i>LSTM-GRNN</i>	0.818	0.805	0.833
<i>FastText</i>	0.841	0.831	0.852
<i>CNN</i>	0.833	0.864	0.806
<i>WESPAD</i>	0.851	0.845	0.858

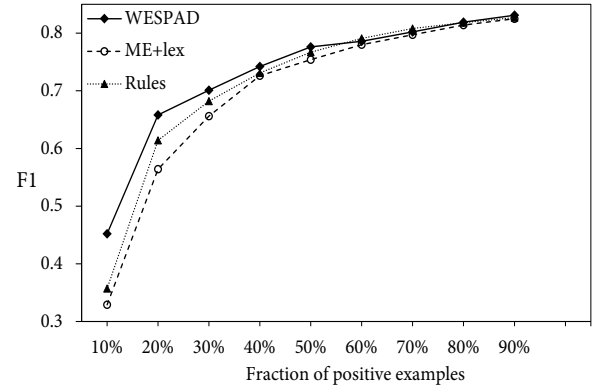
Table 4: F1-measure, precision, and recall in FLU2013 dataset.

Feature set	F1	Precision	Recall
<i>WESPAD</i> (all features)	0.695	0.803	0.628
<i>we_distortion</i>	0.643 (-7.4%)	0.804	0.554
<i>we_partitioning</i>	0.652 (-6.1%)	0.788	0.578
<i>context_next</i>	0.680 (-2.1%)	0.800	0.609
<i>syn_feats</i>	0.682 (-1.8%)	0.800	0.613
<i>context_prev</i>	0.686 (-1.2%)	0.801	0.616
<i>context</i>	0.687 (-1.1%)	0.795	0.620
<i>lex_feats</i>	0.696 (+0.1)	0.782	0.640

Table 5: Feature ablation of WESPAD on PHM2017 dataset.

fractions of the positive tweets used in the training data. The experiment shows that at smaller fractions of available positive tweets (10%-30%), *WESPAD* dramatically outperforms the *ME+lex* baseline, demonstrating that *WESPAD* is able to generalize from fewer positive training examples. *WESPAD* also significantly outperforms *Rules* at small fractions of positive tweets (10%-20%), signifying that the rule based models highly depend on their lexical based counterparts. We also observe that learning from just 20% of the available positive examples, the F1-measure for *ME+lex* model is 0.564, and for *WESPAD* model is 0.658. These F1 values are comparable to the F1 values that these models achieved in PHM2017 dataset, which also contains only 19% of the positive class in the training and test data (on average, across the different disease topics).

In summary, our results show that *WESPAD* is able to outperform the state-of-the-art baselines for both datasets and under variety of settings, and even outperforms a disease-specific classifier in the prominent FLU2013 benchmark dataset. This is striking, as *WESPAD* does not require manual feature engineering, and can be trained with a relatively small number of (positive) training

Figure 4: F1 for *WESPAD*, *Rules*, and *ME+lex* trained on varying subsets of the positive examples (in FLU2013 dataset).

examples which makes *WESPAD* a valuable tool for extending health monitoring over social data to new diseases and conditions.

7 CONCLUSIONS

We presented a new method, *WESPAD*, designed to detect personal health mentions in social data, such as Twitter posts. Unlike previously proposed methods for health classification, our method requires no manual feature engineering, and can be trained on relatively few positive examples of true health mentions. The improvements are due to a new approach to analyzing the representation of the examples in word embedding spaces, allowing *WESPAD* to discover a small number of effective features for classification. Furthermore, *WESPAD* can easily incorporate additional domain knowledge and can be extended to detect new diseases and conditions with relatively little effort.

Our experimental evaluation compares *WESPAD* to a variety of previously proposed methods, including three state-of-the-art deep neural network approaches (LSTM, FastText, and CNN), on both an established benchmark dataset for detecting Flu infection reports, and a new PHM2017 dataset we created, with manual annotations of mentions for six different diseases and conditions. In the majority of the conditions, *WESPAD* exhibits superior overall performance.

By requiring a smaller number of training examples to achieve state-of-the-art performance, *WESPAD* can enable rapid development of domain-specific and robust text classifiers, which could in turn be valuable for tracking emerging diseases and conditions via social media.

REFERENCES

- [1] Charu C. Aggarwal and ChengXiang Zhai. 2012. *A Survey of Text Classification Algorithms*. Springer US, Boston, MA, 163–222.
- [2] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1568–1576.
- [3] David Bamman and Noah A. Smith. 2015. Contextualized Sarcasm Detection on Twitter. In *Proceedings of the Ninth International Conference on Web and Social Media, (ICWSM 2015)*, 2015. 574–577.
- [4] Carmen Banea, Di Chen, Rada Mihalcea, Claire Cardie, and Janyce Wiebe. 2014. SimCompass: Using Deep Learning Word Embeddings to Assess Cross-level Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*. 560–565.
- [5] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, (ICWSM 2010)*, 2010. 10–17.
- [6] Lauren E. Charles-Smith, Tera L. Reynolds, Mark A. Cameron, Mike Conway, Eric H. Y. Lau, Jennifer M. Olsen, Julie A. Pavlin, Mika Shigematsu, Laura C. Streichert, Katie J. Suda, and Courtney D. Corley. 2015. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. *PLOS ONE* 10, 10 (10 2015), 1–20.
- [7] 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 (11 2010), 1–13.
- [8] Munmun De Choudhury. 2015. Anorexia on Tumblr: A Characterization Study. In *Proceedings of the 5th International Conference on Digital Health 2015, Florence, Italy, May 18-20, 2015*. 43–50.
- [9] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, (ICWSM 2013)*, 2013. 1–10.
- [10] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. 2098–2110.
- [11] Aaron M. Cohen and William R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics* 6, 1 (2005), 57–71.
- [12] Hongying Dai, Brian R. Lee, and Jianqiang Hao. 2017. Predicting Asthma Prevalence by Linking Social Media Data and Traditional Surveys. *The ANNALS of the American Academy of Political and Social Science* 669, 1 (2017), 75–92.
- [13] Raminta Daniulaityte, Lu Chen, R. Francois Lamy, G. Robert Carlson, Krishnaprasad Thirunarayan, and Amit Sheth. 2016. “When ‘Bad’ is ‘Good’”: Identifying Personal Communication and Sentiment in Drug-Related Tweets. *JMIR Public Health Surveill* 2, 2 (24 Oct 2016), e162.
- [14] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232.
- [15] Edouard Grave, Tomas Mikolov, Armand Joulin, and Piotr Bojanowski. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, (EACL 2017)*, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers. 427–431.
- [16] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. 873–882.
- [17] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016. 1638–1643.
- [18] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark James Carman. 2016. Are Word Embedding-based Features Useful for Sarcasm Detection?. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2016)*, 2016. 1006–1011.
- [19] Tom Kenter and Maarten de Rijke. 2015. Short Text Similarity with Word Embeddings. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*. 1411–1420.
- [20] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2014)*. 1746–1751.
- [21] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archana Bhatia, Chris Dyer, and Noah A. Smith. 2014. A Dependency Parser for Tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2014)*. 1001–1012.
- [22] Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, (NAACL 2013)*. 789–795.
- [23] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323, 5915 (2009), 721–723.
- [24] Moreno MA, Christakis DA, Egan KG, Brockman LN, and Becker T. 2012. Associations between displayed alcohol references on facebook and problem drinking among college students. *Archives of Pediatrics and Adolescent Medicine* 166, 2 (2012), 157–163.
- [25] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.
- [26] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'05)*. Springer-Verlag, Berlin, Heidelberg, 301–311.
- [27] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. (2002).
- [28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119.
- [29] Tom M. Mitchell. 1997. *Machine learning*. McGraw-Hill Boston, MA.
- [30] Yishai Ofran, Ora Paltiel, Dan Pelleg, Jacob M. Rowe, and Elad Yom-Tov. 2012. Patterns of Information-Seeking for Cancer on the Internet: An Analysis of Real World Data. *PLOS ONE* 7, 9 (09 2012), 1–7.
- [31] Alexandra Olteanu, Emre Kiciman, and Carlos Castillo. 2018. A Critical Review of Online Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 785–786.
- [32] Michael J. Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM 2011, Barcelona, Catalonia, Spain, July 17-21, 2011*. 265–272.
- [33] Michael J. Paul and Mark Dredze. 2017. *Social Monitoring for Public Health*. Morgan & Claypool Publishers.
- [34] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*. 1532–1543.
- [35] Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. 2011. Identifying Health-Related Topics on Twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 18–25.
- [36] Victor M. Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and Jose Luis Oliveira. 2014. Twitter: A Good Place to Detect Health Conditions. *PLOS ONE* 9, 1 (01 2014), 1–11.
- [37] Marcel Salathe and Shashank Khandelwal. 2011. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLOS Computational Biology* 7, 10 (10 2011), 1–7.
- [38] Erich Schubert, Alexander Koos, Tobias Emrich, Andreas Züfle, Klaus Arthur Schmid, and Arthur Zimek. 2015. A Framework for Clustering Uncertain Data. *Proc. VLDB Endowment* 8, 12 (Aug. 2015), 1976–1979.
- [39] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, (NIPS 2013)*, 2013. 926–934.
- [40] Damiano Spina, Julio Gonzalo, and Enrique Amigó. 2013. Discovering Filter Keywords for Company Name Disambiguation in Twitter. *Expert Syst. Appl.* 40, 12 (Sept. 2013), 4986–5003.
- [41] Duyu Tang, Bing Qin, and Ting Liu. 2015. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2015)*, 2015. 1422–1432.
- [42] Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17)*. ACM, New York, NY, USA, 1191–1198.
- [43] Zhiyun Yin, Daniel Fabbri, Trent S. Rosenbloom, and Bradley Malin. 2015. A Scalable Framework to Detect Personal Health Mentions on Twitter. *Journal of Medical Internet Research* 17, 6 (05 Jun 2015), e138.
- [44] Mo Yu, Tiejun Zhao, Daxiang Dong, Hao Tian, and Dianhai Yu. 2013. Compound Embedding Features for Semi-supervised Learning. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, (NAACL 2013)*, 2013. 563–568.