A Hierarchical Learning Model for Extracting Public Health Data from Social Media

Elahm Rastegari
University of Nebraska at Omaha
erastegari@unomaha.edu

Sasan Azizian
University of Nebraska at Omaha
Sazizian@unomaha.edu

Hesham H. Ali
University of Nebraska at Omaha
hali@unomaha.edu

Abstract
In decision-making processes, particularly in the healthcare domain, each relevant piece of information is important. This is particularly important when it comes to the health conditions for them there remains a high degree of non-determinism regarding treatment approaches. Online social media are places in which people feel free to share their opinions about numerous topics, including public health issues and how individuals have perceived the efficacy of different types of treatments associated with diseases. social media could represent a secondary source that can be used as a supplement to other data sources. This would allow individuals as well as healthcare providers to gain insight related to public health from different angels. In this study, we construct a hierarchical learning model based on Twitter data that can extract valuable knowledge associated with public health. Back pain was selected for our case study to demonstrate how the proposed model works.

Keywords
Twitter, public health, sentiment Analysis.

Introduction
In the modern societies, data-driven decision making has been slowly replacing ad-hoc approaches in addressing key issues. In the healthcare domain, the development of effective healthcare systems by health care providers can be significantly enhanced by taking into consideration a full understanding of consumers' concerns and ideas. On the other hand, making appropriate health-related decisions by individuals requires more comprehensive domain-related knowledge. Ideas and opinions provided by various types of individuals through social media represent a good source of information for individuals in making health-related decisions. Such opinions are also valuable for healthcare providers in gaining a better understanding of how various options for treatments have been received by patients. When various treatment options are presented for patients, they rely on data gathered from various sources and experiences of other patients in making health-related decisions (Cambria, Schuller, Xia, & Havasi, 2013).

People use social networks to seek social awareness and acquire health information (Park, McDonald, & Cha, 2013). Twitter is one of the most popular social networks and micro-blogging services, enabling millions of users to communicate ideas by sending and receiving tweets which are messages limited to 140 characters (Teutle, 2010). Recently, analyzing Twitter network data has gained a growing attention. Twitter data have been used in flu epidemic studies and flu trend prediction (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Aramaki, Maskawa, & Morita, 2011). However, nothing has been reported regarding the
A Hierarchical Learning Model for Extracting Public Health Data from Social Media

analysis of Twitter messages to identify treatments mostly taken by Twitter users for a specific disease and to extract Twitter users’ ideas about various disease-specific treatments. We propose a general hierarchical learning model. The proposed model will be used to answer the following questions:

How is it possible to understand Twitter users’ overall perception of the efficacy of treatment approaches for a specific disease?

Is it possible to identify mostly taken treatments associated with a disease by Twitter users?

What is the role of the Twitter network in people’s communication when they post disease-related messages?

What is professional parties’ role in public health information communication in Twitter, considering the number of prevention-related messages and treatment-related messages?

In order to answer the first question, this model classifies messages into individual and professional categories. Then, irrelevant tweets are removed and messages containing treatments’ name are extracted. Finally, this model classifies treatment-related messages posted by individuals into three categories of neutral, positive, and negative messages. The outcome of this step can be used in integration with other sources of information and help physicians and people to make a better decision.

Sentiment analysis, also called opinion mining, is a type of text analysis with the goal of supporting decision making process by measuring how negatively or positively an entity is regarded (Pang & Lee, 2008, Cambria, E. et.al, 2010 , Kim & Hovy, 2004) An increasing number of people are using Twitter to post their opinion. As a result, tweets sentiment analysis is an effective way of evaluating public opinion for public health studies. In general, Twitter sentiment analysis has been done in many studies (Rosenthal, Ritter, Nakov, & Stoyanov, 2014, Pak & Paroubek, 2010, Kouloumpis, Wilson, & Moore, 2011). However, this study is the first major attempt in the domain of treatment sentiment analysis using the Twitter data. In this study, we will identify specific useful features to differentiate positive, negative and neutral messages about various treatments associated with various types of diseases.

In order to answer the second question, we review the literature for a specific disease and then search for those treatments in our Twitter dataset. Assuming the frequency of messages related to each treatment can be correlated to the frequency a treatment has been taken, we will calculate the frequency of messages related to each treatment and then compare the results with data from the PatientsLikeMe website. PatientslikeMe is an online community in which patients share their experience and exchange information. To answer the third question, we will categorize messages into two categories: the messages asking questions and other messages (either posted to give useful information or posted to share users’ status). Then we will calculate the frequency of messages in each category. The answer to the last question will be provided by investigating whether professionals provide people with more information regarding the causes of diseases and preventive ways or they do post more information regarding treatments.

Back pain is one of the costly health issues for which a very high socioeconomic cost has been reported in the UK and US and it is in accordance with findings in other countries (Deyo et al., 2014; Maniadakis & Gray, 2000). Substantial personal and financial burden along with the lack of a consensus about various back pain treatments among physicians (Chou et al., 2009, Deyo, R. A., 2014) makes people eager to look for their peer’s opinion regarding back pain and get involved in online health discussions through social networks. In this study, our focus will be specifically on Twitter messages related to back pain to demonstrate how our model works and what other public health information can be learned from Twitter.

Our proposed model will be explained in the next section which will be followed by a section on data collection and preprocessing. Then each step in the proposed model will be described in details in the next following sections

Method

Analyzing Twitter messages to understand people’s opinions regarding treatments is a challenging task and involves several essential steps. The diagram in Figure 1 shows our proposed hierarchical model.

1 https://www.patientslikeme.com/
Data Collection and Preprocessing

Most of the Twitter users' profiles are publicly available (Mislove et al., 2011). We used NodeXL (Sameh, 2013) to crawl the Twitter network and collect tweets. The data was collected from June 16th, 2016 till September 20th, 2016, using “back pain” as our search keywords. After removing duplicates and tweets in languages other than English, totally 11689 tweets were collected. After collecting data, retweets were removed because they usually refer to the news and advertisement and rarely describe individuals’ opinion (Sadilek, et al., 2012). 7562 messages remained after removing the retweets. Preprocessing of tweets, is a necessary step affecting accuracy, and has been considered prior to sentiment analysis of Twitter messages by other researchers (Pak & Paroubek, 2010, Kouloumpis, Wilson, & Moore, 2011).

In this study, the following steps were considered in the preprocessing phase:

1) Substitution of URLs and usernames: usernames were substituted after they were used in differentiating messages posted by individuals from those posted by professionals.
2) Normalization: normalization was done by converting upper case to lower case and replacing letters and characters happening more than two times consecutively with two occurrences. Short forms were substituted with full forms.
3) Tokenization: we split the texts by spaces. Numbers and stop words (articles) were removed.
4) N-grams: a set of unigrams, bigrams, and trigrams were made.

Linguistic Inquiry and Word Count (LAWC) is a validated toolkit that is able to reveal the language used in various domains by categorizing the words into personal concerns processes, spoken processes, psychological processes, and linguistic processes (Pennebaker, et al., 2015).

Individual vs Professional

In order to obtain a clear idea of what individuals think about various treatment approaches, Twitter messages should be categorized into two classes: messages posted by individuals and messages posted by professionals. We used a supervised learning approach. “Individuals” are defined as users who have a personal account and “professionals” are defined as users associated with companies, news channels or health-related advertising websites.

Author and features representing emotions have been reported to be useful in differentiating corporate twitters from personal twitters (Sriram, et al., 2010). In this study, along with usernames, the following...
attributes, extracted using LIWC2015, were used as our classifier features: emotion category features, word count, words per sentence, personal pronouns, netspeak and informal. We selected these features because there was a significant difference between the mean value of these features for individuals and professionals in the training data set (based on t-test). The mean values of emotion features, personal pronouns, netspeak and informal features in the individual group were significantly higher than of the professional group. A logistic linear classifier was performed. The average accuracy of the model, using 5-fold cross validation technique was 90.5%. Precision and recall were 91.2% and 90.4%, respectively. Out of 7562 messages, 2908 were classified as messages posted by “individuals” and 4654 messages were classified as “professionals”.

Relevance of Messages

In our case study, “unrelated tweets”, are considered tweets such as “I wish we could get back together. I am in pain”, containing both words “back” and “pain”, but not meaning “back pain”. A supervised learning technique was applied to remove unrelated tweets. We performed a series of t-tests for LIWC features between two sets (relavant and irrelevant messages labeled as our training data set). Based on Sig values, we observed that almost all LIWC features were distributed evenly in two sets. Therefore, we did not use the LIWC features as our classifier features. Instead, we used unigrams and bigrams. The presence of each unigram or bigram was considered as one binary feature. Using SVM with linear kernel, out of 2908 messages submitted by individuals, 1882 messages were classified as related with 87% accuracy, 83.5% precision and 88% recall. From professional data set, 3868 messages were classified as related.

Role of Twitter in People’s Communications about any Specific Disease

At this point, we can identify the role of Twitter network in the people’s communications regarding any specific disease by answering the following question: Have Twitter users sought information by directly asking questions or they post/read information when it comes to disease-related communication?

To answer this question, we decided to apply learning technique to identify between messages asking questions and other messages. However, in the training phase, two reviewers realized that only 4 messages out of 400 messages were asking a question. Only 5 more questioning messages were found in another set of 400 messages. This class imbalance is an indicator of having very less questioning messages compared to the other messages. This means that people do not use Twitter to directly ask questions about back pain. If the goal was to identify questioning messages, over-sampling or under-sampling (Chawla, et al., 2002) could be used to solve the class imbalance problem.

Analyzing Treatment-Related Messages

In order to analyze messages related to disease-specific treatments, first we need to find as many possible treatments as exist for a specific disease. Then we can extract treatment-related messages and analyze them.

Finding Various Existing Treatments for a Disease

A comprehensive literature review regarding existing treatments for a specific disease is necessary. In the case of back pain, based on the literature, there are lots of treatment for lower back pain. We considered as much various treatment as we found in the literature. Since tweeters’ language can be different from scientific language and everyday language, various phrases or words that people might use in calling a treatment should be taken into consideration. In Table 2, various treatments and different words and phrases people used in Twitter messages regarding each treatment approach can be seen. All messages about each treatment were extracted from both individual and professional datasets. The total number of messages individuals and professionals posted regarding all types of treatments were 571 and 2417, respectively. We also looked at PatientsLikeMe website for the number of people with back pain who are taking or has taken each of the treatments listed in Table 2. We had a total number of 590 reports regarding these treatment approaches.
A Hierarchical Learning Model for Extracting Public Health Data from Social Media

Treatments | Search words and phrases
---|---
Physical therapy (Harvey et al., 2003) | Physical therapy
Chiropractic therapy (Harvey et al., 2003) | Chiropractic
Exercise/yoga (Khalil et al., 1992; Sherman, et al., 2005) | Exercise, exercises, stretching, stretch, walking, walk, swimming, yoga
Massage therapy (Van Tulder, et al., 1997) | Massage
Drugs (e.g., gabapentin) (Chaparro et al., 2013, Van Tulder, et al., 1997) | Anti-inflammatory drug, anti inflammatory drugs, pain killer, pain killers, pain meds, anti-inflammatory drug, anti-inflammatory drugs, advil, gabapentin, hydrocodone, flexeril, ibuprofen, acetaminophen, cyclobenzaprine, tramadol, muscle relaxer, hydrocodone-acetaminophen, oxycodone, opioid
Heat/ cold therapy (French, et al., 2006) | Heat, heat pack, heat pad, heat therapy, heating pad Ice, ice pack, temperature, ice pack, ice therapy, cold pack, cold
Spinal injection (Staal, et al., 2013) | Epidural, injection, epidural injection, steroid injection
Placebo (Chaparro et al., 2013) | Placebo
Acupuncture (Lee et al., 2013) | Acupuncture
Back brace (Willner, 1985) | Back brace, brace
Osteopathy (Harvey et al., 2003) | Osteopathy

Table 2. Back Pain Treatments

In Figure 2, we can see the number of the messages associated with each treatment and posted by individuals in the left side. The number of the people who are using or have used each treatment associated with back pain, based on reports in PatientsLikeMe, is shown in Figure 2 in the right side. As it can be seen in Figure 2, “drugs”, “physical therapy”, “chiropractic”, and “massage” are the four top most categories of treatments reported by patients in the PatientsLikeMe website and they are exactly the four top most categories of treatments individuals in Twitter had conversation about. As it is observed in Figure 2, these four categories are exactly in the same order in the PatientsLikeMe and individual’s plots.

Treatment-related messages posted by professionals were analyzed. Figure 3 shows the number of the messages associated with each treatment and posted by professionals. As it is shown in Figure 3, “exercise/yoga”, “massage”, “chiropractic”, and “physical therapy” are the top most categories. The order of the categories identified using messages posted by professionals are different from that of both individuals and PatientsLikeMe. It may be due to the fact that a great number of messages posted by professionals are used for advertisement.

Do Professionals post messages about preventive approaches?

Almost 80% of related messages posted by professionals were found to be treatment-related messages. This means that less than 20% of messages posted by professionals were about prevention. Assuming that professionals might have posted messages about “exercise category” and “yoga” as preventive approaches for back pain, we did a simple search in the associated data sets to find the number of prevention-related messages: Messages included phrases such as “prevent back pain”, “preventing back pain”, “avoid back pain”, and “avoiding back pain” were counted, leading to 8% increase in the percentage of prevention-related messages. This result indicates that at least in the case of back pain, people have been more provided with more treatment-related information rather than prevention-related information.
Mining Users’ Opinion Regarding Treatments

Several subtasks have been identified by previous studies regarding opinion mining, including 1) determining whether a message is objective or subjective (Sameh, A., 2013, Rosenthal, et al., 2014), 2) Determining whether a subjective message expresses a negative or a positive opinion (Rosenthal, et al., 2014), and 3) deciding on the strength of the message polarity (Rajan & Victor, 2014, Rosenthal, et al., 2014). Two first subtasks were considered in our work.

Figure 2. Number of messages posted by individuals in Twitter (in the left) and patients in PatientsLikeMe (in the right) regarding back pain treatments

Subjective vs Neutral Messages

Part of Speech (POS) features have been used in classifying Twitter messages into subjective and objective categories (Rosenthal, et al., 2014, Pak & Paroubek, 2010). The most frequent role of adverbs and adjectives is to carry opinionated contents (Pak & Paroubek, 2010). A series of t-tests was used to find features that could discriminate between subjective and neutral messages. The number of adverbs and adjectives was considered as one of the discriminating features. Subjective messages tend to have more comparing words (e.g., as good as) and fewer personal pronoun compared to objective messages, while authors of objective messages use more personal pronouns, and fewer adverbs and adjectives compared to the authors of subjective messages. The presence of verbs such as help, ease, eliminate, decrease, reduce, relieve, heal, recover, increase, and feel, and these verbs’ variations and negative forms was an indicator of having a subjective message. Therefore, unigrams and bigrams were included in the feature space. Unigrams and bigrams representing the names of the treatments were excluded from our feature space.

Non-binary features (POS), were scaled to [0, 1]. We tried linear kernel SVM classifier with two combinations of features: 1) POS and unigrams, and 2) POS, unigrams, and bigrams. Taking Bigrams into
consideration increased the classifier accuracy from 82.5% to 89.1%. From 571 treatment-related messages posted by individuals, 201 messages were classified as subjective messages.

**More Effective/Less Effective**

We had only a very small dataset of subjective tweets regarding treatments for back pain and we could manually label them as positive or negative. In general, however, this step becomes more and more essential when it comes to analysis of a huge dataset. Since false positives and false negatives should not be traded off against each other in this domain, achieving high accuracy, precision and recall is very important. One possible approach is to learn a linear support vector machine (SVM) classifier $C_{\text{final}}$ and consistently keep high precision and recall by optimizing the area under the ROC curve (Joachims, 2005). Cascade learning approach, used by Sadilek and his colleagues, performed well regarding both precision and recall in classifying tweets associated with two categories (Sadilek, Kautz, & Silenzio, 2012). We applied cascade learning approach in which one part of our dataset and two other datasets collected using positive and negative emoticons were used as our training dataset. Using emoticons from Twitter for sentiment analysis has been reported as one of the successful approaches for collecting a corpus (Pak & Paroubek, 2010, Kouloumpis, et al., 2011). We used the positive and negative emoticons along with the name of each treatment to collect the corpus.

One thousands messages collected using positive emoticons and one thousands negative messages collected using negative emoticons were selected, merged together and called PosNegCorp. First, we trained two helpers SVM $C_{\text{Pos}}$ and $C_{\text{Neg}}$ on a dataset of 80 tweets from our dataset. Each of the messages was labeled as “positive” or “negative” by three reviewers and label with the majority vote was selected. $C_{\text{Pos}}$ was highly penalized if a negative tweet was mistakenly labeled as positive, while $C_{\text{Neg}}$ was highly penalized if a positive tweet was labeled as negative. After training, $C_{\text{Pos}}$ and $C_{\text{Neg}}$ were both used to label PosNegCorp data set. In order to reduce the noise, we only selected 15% of the messages that were positive with high probability and 15% of messages that were negative with high probability. A diagram of our training process is shown in Figure 4. All unigrams and bigrams were used as our classifier features. Applying $C_{\text{final}}$ on a held-out set we gained 83% precision and 84.5 % recall.

<table>
<thead>
<tr>
<th>Positive Features</th>
<th>Negative Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feeling better</td>
<td>Not provide reliefe</td>
</tr>
<tr>
<td>Relief</td>
<td>Not feeling better</td>
</tr>
<tr>
<td>help</td>
<td>Not helpful</td>
</tr>
<tr>
<td>Alleviate</td>
<td>Not help</td>
</tr>
<tr>
<td>Ease</td>
<td>Not effective</td>
</tr>
<tr>
<td>Relieve</td>
<td>Not good</td>
</tr>
<tr>
<td>effective</td>
<td>Ineffective</td>
</tr>
<tr>
<td>Eliminated</td>
<td>Increased pain</td>
</tr>
</tbody>
</table>

Table 3. Examples of positive and negative features found from our classifier
A Hierarchical Learning Model for Extracting Public Health Data from Social Media

Table 4. Examples of neutral, positive, and negative tweets

Table 3 shows a list of significant features found after using cascade learning. Table 4 lists examples of neutral, positive, and negative messages found. The results of this polarity detection regarding various treatments can be seen in Table 5. In the PatientsLikeMe website, patients rated the effectiveness of each treatment. We did not compare our results with the ratings provided there because here we have only two rating levels, while in PatientsLikeMe, patients rated each treatment on a 5-level basis.

Table 5. Number of neutral, positive, and negative messages posted in Twitter about each treatment

Discussion

In this study, we proposed a hierarchical learning approach to analyze Twitter data in the domain of public health and showed how the model operates using a case study. We used Twitter data and demonstrated that it represents a viable data source in mining opinion regarding various treatments. However, due to the potential variability in accuracy and data quality, we are looking at Twitter data as a supplementary data source that can be integrated with other clinical sources to allow us to look at the health-related problems from multiple angles.

Many studies tried to identify the most powerful features and classifiers in differentiating neutral, positive and negative tweets. However, most of the studies have not focused on treatment-related sentiment
analysis. In this study, our introduced model differentiated between messages posted by individuals and professionals. After filtering out irrelevant messages, a sentiment analysis was conducted. However, this study has some limitations. Data was collected over a period of three months, out of which only 5.4% were treatment-related tweets and almost half were subjective. To further evaluate the accuracy of the results obtained by this approach, a larger data set is needed. It would also be critical to include treatment-related messages associated with multiple conditions in addition to back pain. This will be part of our future studies.

A comprehensive research needs to be conducted to rank sentiment analysis techniques, recognize the best features or combination of features, and identify classifiers with the highest accuracy, precision, and recall. In our future work, we plan to examine various combinations of features and apply various classifiers reported in the literature to find the best classifiers for treatment-specific sentiment analysis. Our plan is to apply various classification techniques on various training sets related to public health. Training sets will also be collected using different methods such as emoticons-based dataset, hashtags dataset or manually labeled data to see which classifier perform better with which training set.

Conclusion

The proposed work represents one of the early attempts in twitter data mining to understand how people perceive the efficacy of specific-disease treatment approaches. We developed a model with a hierarchical learning approach that identifies the role of Twitter in the communications professionals and individuals have regarding public health, specifically communications related to a particular disease and its associated treatments. We found that individuals use Twitter to post information/status more than posting messages to ask questions. Our results also indicate that in Twitter, healthcare providers, healthcare centers, news channels and advertising companies provide more information related to treatment rather than information related to prevention. This work also shows that Twitter could be a good source for public health data. Our model and results based on our data set demonstrate that whatever treatments people are taking for back pain in the real life could be found in the Twitter users’ conversations. Furthermore, as a validation point, we found that there is a correlation between treatment usage reported by patients with back pain in the PatientsLikeMe website and the percentage of messages posted about the treatment in Twitter Network. One of the main contributions of this work is that the introduced model is not limited to one condition or disease and it could be used to collect similar useful information associated with other diseases or conditions. Furthermore, this work was able to identify discriminating features for sentiment analysis of treatment-related messages.

REFERENCES


Park, M., McDonald, D. W., & Cha, M. 2013. "Perception Differences between the Depressed and Non-Depressed Users in Twitter". In ICWSM.


