Natural Language Processing for Health and Social Media
Authors: Mark Dredze, Michael J. Paul

Intro

Social media, such as Twitter, has shown great potential to analyze real world events, such as politics, product sentiment and natural disasters. In recent years, social media has emerged in the health community, particularly in public health, as a revolutionary data source for a wide range of problems. Vast amounts of naturalistic population data can be collected through social media much faster and at lower cost than through traditional data sources such as surveys. Additionally, social media provides novel data previously unavailable to researchers. These advantages allow for the rapid formulation and evaluation of novel hypotheses, aiding decisions about how best to spend limited traditional data collection resources.

While data from social media is plentiful, it can be difficult to utilize. Most health applications require querying the data as if it were available as a structured database (e.g., “how many messages about flu infection on each day?”), while the data arrive in the form of unstructured text. The underlying data attributes (e.g., “does this message indicate a flu infection?”) are not directly available and can only be inferred. This is the key challenge of mining health information and trends from raw text: the structure must be inferred from unstructured data, but only automated methods are viable at scale.

A basic information retrieval approach to this problem is to query for messages containing relevant keywords, such as “flu” or “fever.” A limitation of this approach is that it ignores the context of these words. For example, the word “flu” could indicate a person is sick, or it might just be an acknowledgement of news articles about an ongoing flu season. The word “sick” itself has many colloquial meanings that are prevalent on social media beyond indicating illness. Methods based on *natural language processing* (NLP), an area of computer science focused on developing algorithms to understand human language, can handle these limitations by making use of richer context. Even when simple keyword querying works well, utilization of more sophisticated NLP algorithms can lead to significant improvements [1]. Additionally, for other applications, NLP provides opportunities otherwise unavailable, such as discovering health issues in social media messages, or trends in sentiment about physicians in doctor reviews.

In this article, we describe our work that demonstrates the potential of NLP for several health applications with social media data.

Discovering Health Issues

When confronted with massive amounts of social media data, it is difficult to manually survey and explore the data, or know *a priori* what to expect, an important step in judging its suitability for various tasks. Searching for specific keywords may miss major topics since the user may not know what to search.

We sought to discover which health issues are commonly discussed in social media using *topic models*, a class of NLP tools that can be used to organize large volumes of text. With minimal human intervention, topic models can automatically discover

prominent themes (“topics”) in text without specific knowledge of the data set. These statistical models assume documents are composed of underlying distributions of topics, where topics in turn are formed as distributions of words. The result of statistical inference is a clustering of words and documents into topics, which can then be reviewed by a user to identify major themes in the data.

For Twitter data, we created the Ailment Topic Aspect Model (ATAM), a specialized topic model that explicitly distinguishes health topics from other topics by incorporating general knowledge of symptoms and treatments as extracted from a health website. When ATAM was applied to a large health Twitter collection -- 1.6 million health-related tweets from 2009-2010 -- it discovered fifteen health topics or ailments, including allergies, aches and pains, dental health, and insomnia [2].

We validated these discovered ailments as a public health resource by measuring their ability to capture real world trends of health lifestyle factors as measured by the U.S. Behavioral Risk Factor Surveillance System (BRFSS), a large annual survey of American adults (Figure 1). Our analysis showed several promising correlations, such as tweets about cancer and serious illness being positively correlated with the prevalence of smoking across U.S. states (0.56), and tweets indicating obesity negatively correlated with rates of exercise (-0.44). As a control, we include two ailments that we expect to have no correlation (0.02): bacterial infection (ATAM) and asthma (BRFSS). These findings demonstrate that health topics discussed in tweets closely match behavioral patterns in the real world. This NLP-driven analysis highlights health areas that are supported by the data, leading to new directions of study.

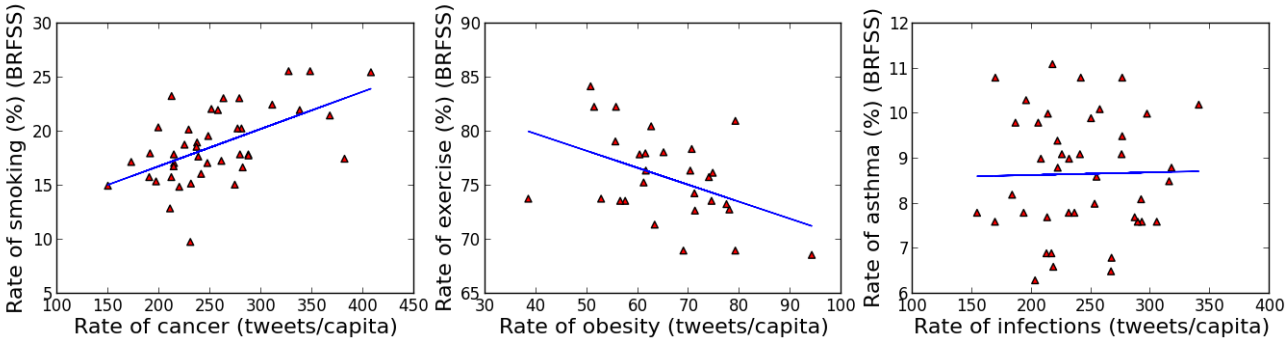

Figure 1: Scatterplots of the volume of tweets per U.S. state for various ailments which are positively, negatively, and not correlated with government survey data.

Deeper Linguistic Analysis for Influenza Surveillance

While ATAM provided data exploration capabilities, it only considered word usage, ignoring the rich linguistic features that are the core of NLP algorithms. For the task of influenza tracking, which requires algorithms that determine the intention of each tweet, we utilized supervised machine learning with rich features for deeper language analysis to make subtle distinctions in the texts.

For the task of influenza surveillance, our goal is to measure the current prevalence of influenza as captured by the rate of tweets about influenza. With this particular goal in mind, we wish to incorporate our own knowledge, expectations and observations of the data. We observed that not all tweets containing flu-related keywords expressed an actual infection. Especially during severe seasons with high media attention, many tweets express a concern or awareness of the ongoing flu season even if the user is not personally sick. These types of tweets, which still contain relevant keywords, will lead to incorrect estimates of flu prevalence.

To obtain more accurate influenza measures, we trained supervised classifiers on over 10,000 tweets labeled by humans to determine if a specific tweet reported an influenza infection. We classified tweets as related or unrelated to health, then related or unrelated to flu, and finally descriptive of flu infection or not. These subtle distinctions in language required the construction of a rich set of linguistic features, including longer phrases (n-grams), human-created groups of keywords (for example, concern-related words like “worried” and “scared”), Twitter-related features like URLs, hashtags, user mentions and emoticons, and linguistic features using part-of-speech (e.g. noun, adjective) information associated with each word. For example, when flu is the subject of a sentence it indicates awareness, e.g. “the flu is going around,” while as an object indicates infection, e.g. “sick with the flu.” Flu as an adjective often suggests awareness, e.g. “flu season” or “flu shot.”

Our experiments confirmed that this deeper NLP analysis yielded significant improvements over simpler keyword approaches. We compared our Twitter-based flu estimates to the weekly influenza rates from the U.S. Centers for Disease Control and Prevention (CDC). Figure 2 shows these rates side by side from December 2011 through May 2013, and the two trends have a strong correlation of 0.85, a correlation significantly improved from a keyword baseline, 0.68.

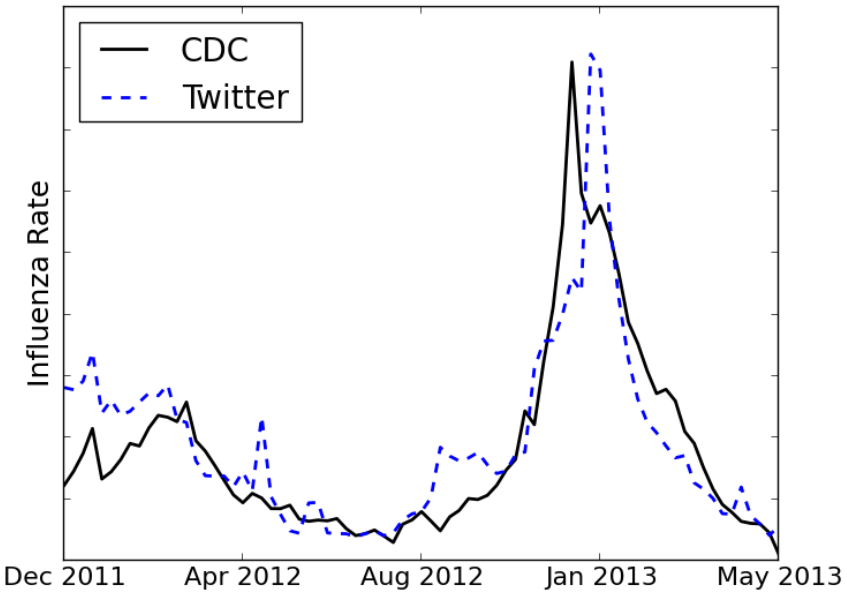

Figure 2: Our Twitter-derived estimate of influenza prevalence in the U.S. over time, along with the rate measured by the U.S. Centers for Disease Control and Prevention (CDC). The two trends have a correlation of 0.85.

Beyond Twitter
Twitter, as general-purpose platform with an enormous user base, is a great source for tasks that require general population monitoring, such as influenza and behavioral risk factor monitoring. However, many health-focused online communities provide more detailed health information [3]. With expansive texts, analyzing these communities requires NLP algorithms. We will briefly describe tasks on two such communities.

RateMDs: We analyzed a set of 50,000 doctor reviews from RateMDs.com, a website where users share reviews of healthcare providers. Using topic models for sentiment analysis -- an NLP task that infers opinions from text -- we inferred the sentiment expressed in reviews for various topics like interpersonal manner, which can be used to analyze patient perceptions of healthcare across the U.S [4].

Drugs Forum: We analyzed a set of 400,000 messages from Drugs-Forum.com, a discussion forum where users anonymously discuss illicit drug activity. Using topic models as a basis for multi-document summarization -- extracting snippets of text that concisely reflect many related posts -- we summarized discussions from many forum messages to reveal specific information on drug use, such as the effects and dosage of drugs [5]. Our summarization system correctly identified the typical dose and common side effects of mephedrone, a new and potentially dangerous drug.

These examples highlight just some of the opportunities for NLP to reveal valuable public health information from social media text.

Looking Forward
When considering social media sites such as Twitter and online health communities, there is no shortage of opportunities for the rapid identification of important health information -- information often unavailable through traditional health resources. Critically, at the core of social media is human language, and NLP technologies are required to mine these data. We have found that while out-of-the-box methods work well, custom solutions can provide valuable improvements for specific tasks. These solutions blend familiarity with social media data and domain expertise to elicit the right health information. An effective NLP approach begins by choosing the right data source -- a site like Twitter for general monitoring of population health effects versus online forums for specialized medical questions -- and continues with a clear formulation of the health question informed by domain expertise. This approach may lead to new NLP algorithms that can go far beyond surface level processing, e.g. keyword filtering, to deliver exciting new data sources for health.

References

[1] A. Lamb, M.J. Paul, and M. Dredze. “Separating fact from fear: Tracking flu infections on Twitter,” in *Proc. NAACL-HLT*, 2013, pp. 789-795.

[2] M.J. Paul and M. Dredze. “You are what you Tweet: Analyzing Twitter for public health,” in *Proc. ICWSM*, 2011, pp. 265-272.

[3] M. Jha and N. Elhadad. “Cancer stage prediction based on patient online discourse,” in *Proc. ACL BioNLP Workshop*, 2010, pp. 64-71.

[4] M.J. Paul, B.C. Wallace, and M. Dredze. “What affects patient (dis)satisfaction? Analyzing online doctor ratings with a joint topic-sentiment model,” in *Proc. AAAI HIAI Workshop*, 2013, pp. 53-58.

[5] M.J. Paul and M. Dredze. “Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models,” in *Proc. NAACL-HLT*, 2013, pp. 168-178.