4-15-2016

Finding Specific, Topic Related Information from a Sea of Social Media Posts

Scott J. Duberstein  
*Wright State University - Main Campus, duberstein.4@wright.edu*

Daniel Asamoah  
*Wright State University - Main Campus, daniel.asamoah@wright.edu*

Derek Doran  
*Wright State University - Main Campus, derek.doran@wright.edu*

Shu Z. Schiller  
*Wright State University - Main Campus, shu.schiller@wright.edu*

Follow this and additional works at: [http://corescholar.libraries.wright.edu/urop_celebration](http://corescholar.libraries.wright.edu/urop_celebration)

Part of the [Arts and Humanities Commons](http://corescholar.libraries.wright.edu/urop_celebration), [Engineering Commons](http://corescholar.libraries.wright.edu/urop_celebration), [Life Sciences Commons](http://corescholar.libraries.wright.edu/urop_celebration), [Medicine and Health Sciences Commons](http://corescholar.libraries.wright.edu/urop_celebration), [Physical Sciences and Mathematics Commons](http://corescholar.libraries.wright.edu/urop_celebration), and the [Social and Behavioral Sciences Commons](http://corescholar.libraries.wright.edu/urop_celebration)

Repository Citation

Finding Specific, Topic Related Information from a Sea of Social Media Posts
Scott Duberstein, WSU Undergraduate Student (Presenter), Daniel Asamoah, Ph.D. (PI), Derek Doran, Ph.D. (Co-PI), Shu Schiller, Ph.D. (Co-PI)

Abstract
As social media continues to become an incredible mode of communication in daily life dealing with the exchange of information, these systems provide authors a platform where they can share their thoughts, feelings, and experiences about a number of topics. Harnessing the information expressed publicly through these modes can be incredibly powerful: public perceptions, signals, and data about a variety of specific topics could be extracted and studied from these posts. However, there is a common trade-off in collecting information about a topic from social media: the more specific the topic, generally, the more challenging it is to extract meaningful information. This is because, at first glance, social media posts are simply too noisy authors post topics that are forced to inject meaning in a short length (140 characters on Twitter). This work presents a novel methodology to overcome this problem. It uses state-of-the-art programming and data storage technologies, stop-word dictionaries, author filters and Twitter bot detectors. Short of evaluating the authenticity of the collected tweets, which will be done in future work through Amazon Mechanical Turk evaluators, we demonstrate how our methodology extracts specific, meaningful tweets about topics related to chronic diseases and medication.

Implementation
During querying the Twitter API tweets are evaluated to determine whether or not they should be stored in a PostgreSQL database. Setting flags at the start of an initial query is paramount to removing bots. Manual reviews are neglected from any results as these make up for a vast amount of tweets made by bots. Before writing to the database the tweets are passed through a stop-word dictionary. The dictionary contains pop culture references in order to choose accounts that are attempting to tell a real story as opposed to ones that are picking on celebrities or politicians who, “forgot to take their meds”. The evaluation process tests if the database has too many instances of Kanye West and Lexapro, then it is safe to assume the tweet is a lyric and remove it safely because it contains no relevant content. This rule applies with accounts that tweet the same material over and over.

Given the high traffic volume of Twitter and the infinite access of timeline information one has with the Twitter API, identifying bots and accounts attempting to sell and promote products, from humans who have had genuine experiences with medication is crucial. Determining who had real information became a key sticking point.

Content accounts, accounts that tweet only on a specific topic, are accounts that do both evaluation and gathering of “supported” medical advice without searching with the Twitter API. They are actively retweeting, tweeting advice, and sharing links with “useful information.”

Testing
Once implementation was complete and queries were returning positive results, the real human side began to show. Even at its intended core, the “real” content, the stories that mattered are not necessarily taking place on content accounts or via hashtags, but via @ replies. People are having more intimate conversations with one another through @ replies. @ replies make up more than 45 percent of 400 tweets related to just prescription drugs.

Future Goals
1. Incorporate the use of crowd sourcing with Amazon Mechanical Turk to determine sentiment and truth to each tweet's validity.
2. Use natural language processing to better make sense of tweets that are pulled from the same hashtag. Where some may feel #depressed about exams others may actually be talking about what it actually feels like to be depressed.
3. Identify social networks that are actively supporting each other and exchanging information through the use of ego networks.
4. Analyze sentiment over time given a tweet that states success or failure in treatment with a particular drug or technique with coping using both Amazon Mechanical Turk and algorithms.

Acknowledgements