Background

As rates of chronic disease increase in the United States and around the world, it is important for researchers and policymakers to understand how people manage long-term health conditions. Social media posts provide a rich source of first-person lifestyle information, but it is difficult to extract meaningful information about the health of a population from these posts. People frequently turn to their social media platforms to discuss symptoms from various ailments. The Centers of Disease Control (CDC) wanted to understand if they could mine social media data to understand population health. They needed a data science firm to complete this task.

Mosaic Data Science, a leading analytics consulting firm, assembled a team of data scientists and software engineers to build a proof-of-concept natural language processing (NLP) tool for this purpose. Over just a few short months, the team scraped two health-based online forums, HealingWell.org and Mothering.com, and integrated the posts into a web-based exploratory data analysis tool called the Mosaic Social Media Context Awareness Platform. The platform was designed to assess the feasibility of grouping social media users into chronic disease cohorts based on users' characteristics. If successful, researchers could mine data from a single disease cohort to learn about recent trends in adherence to different treatments, behaviors associated with better and worse outcomes, and the day-to-day symptoms and concerns of people with particular conditions.

Analysis

Reading individuals' social media posts can be interesting, but to be useful for larger-scale research, a data science consultant must be able to combine information from many people and to compare the experiences of different demographic groups. The difficulty with social media is people do not explicitly say who they are, where they live and their demographic characteristics. To fill this gap, Mosaic's data scientists used machine learning to predict demographic characteristics of the writers of social media posts based on patterns identified in other posts. The process goes like this:

1. First, all posts are pre-processed using NLP techniques to extract the root words in each post.
2. The user decides on a category by which to separate posts (e.g., gender).
3. The user provides specific text strings that identify writers of social media posts as belonging to particular groups. For example, "I am a 23-year-old man" might be associated with the group "male."
4. A pattern matching script takes all of the text patterns, and labels any social media posts with the matching strings with the specified group label (e.g., male or female).
5. A support vector classification model is trained on the other words in the labeled posts (i.e., everything except the specified patterns) to learn words associated with group membership.
6. The classifier predicts the group to which each of the unlabeled posts is likely to belong.
7. All predicted and pattern-matched posts are assigned a group label, and the category can now be used in exploratory data analysis.

For this prototype platform, the linear support vector machine used single words to predict group membership. An example of the classification strategy is shown for the gender category in Figure 1. The chart shows the features that were the strongest positive (blue) and negative (red) predictors of male social media post writers.

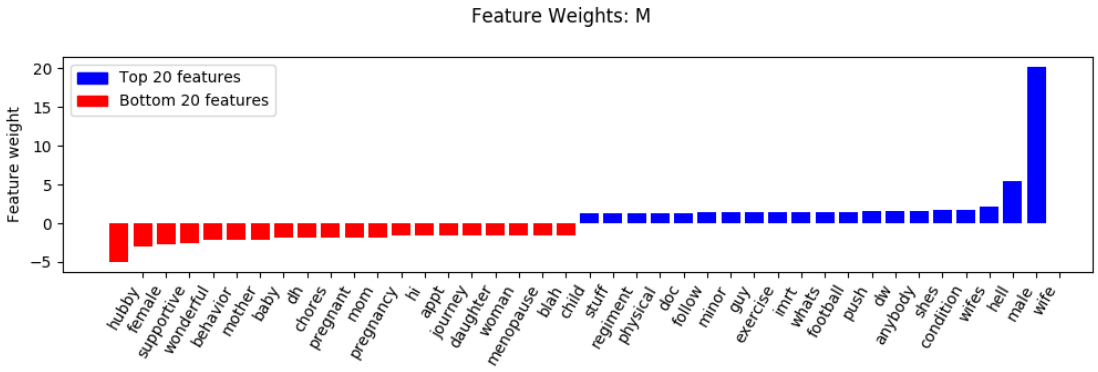

Figure 1. Top and bottom 20 words associated with the male gender group. Male social media users were very likely to use the words shown under the blue bars in the chart, and they were unlikely to use the words shown under the red bars.

The Mosaic Social Media Context Awareness Platform is set up so that any user can create a new category with any group of interest, thereby maintaining the flexibility to search a variety of topics within the posts. For this prototype design, Mosaic decided to enable three types of visualizations based on the predicted categories and other categories determined through metadata from the posts (e.g., chronic disease type). The prototype also allows posts to be filtered by additional keywords, and to visualize only certain groups within a category.

An example of one analysis type, the Cross Correlation Analysis, is shown in Figure 2. The chart shows the number of posts in each of two chronic disease groups, Chronic Pain and Mental Health, by social media users in different age groups. The results show that while mental health is a primary concern among young adults, social media users in their 50's and early 60's are at least as concerned with chronic pain. This type of insight, if supported by further research, could help to direct health outreach to the particular conditions most problematic among different age groups.

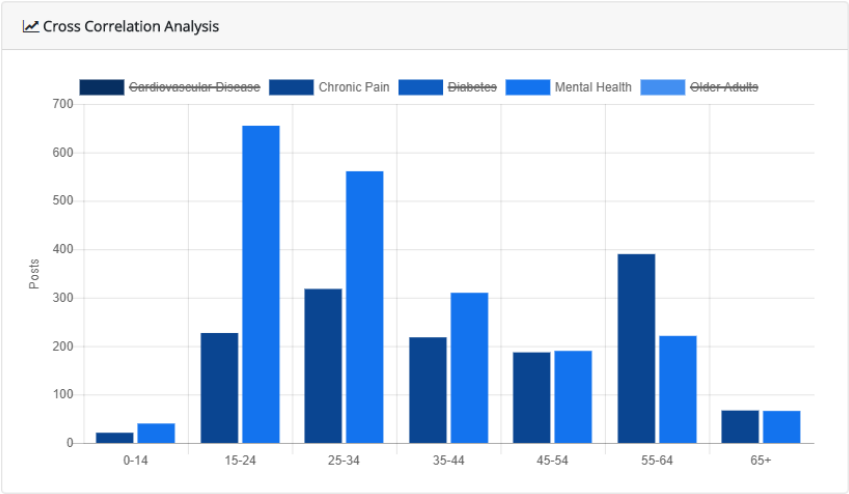

Figure 2. The number of posts about Chronic Pain and Mental Health by social media users of different ages. Older adults are much more likely to discuss pain-related issues on these particular health forums, whereas younger adults are more likely to discuss mental health concerns.

Results

Mosaic Data Science designed and prototyped a web platform that facilitates interactive analysis of social media data related to chronic health conditions and disease management. This proof-of-concept effort set out to test the feasibility of using natural language processing to extract meaning from social media posts and associated metadata for exploratory public health research. A successful prototype was built that establishes the infrastructure needed to support a system of post-aggregation and interactive visualization. Mosaic used semi-supervised machine learning algorithms to characterize posts grouped by particular word patterns that can then be applied to predict characteristics of additional social media posts. This modular platform enables initial exploration of the web application while providing the flexibility to add analytic modalities and enhanced social media data mining and prediction algorithms in the future.

The prototype platform can be accessed at this website: https://cdctool.mosaicatm.com

For More Information

Want to learn more? Please contact info@mosaicdatascience.com