Mental Illness and Media Stigmatization Analytics

Eric Goplerud, Ph.D.

In the immediate aftermath of the shooting at Sandy Hook Elementary that left 20 children and seven adults dead, there was much speculation in the press about the mental health of the shooter.

In response to the starkly stigmatizing language used in media portrayals of the shooter and public debates about restrictions on gun ownership, The Associated Press (AP), on March 7, 2013, issued new guidance in the AP Stylebook concerning appropriate, non-stigmatizing description of mental illness and people with mental illness for print, online, and broadcast media. The Stylebook entry begins: "Do not describe an individual as mentally ill unless it is clearly pertinent to a story and the diagnosis is properly sourced." The AP Stylebook is considered the newspaper industry standard, and is also used by broadcasters, magazines, and public relations firms (http://www.apstylebook.com/).

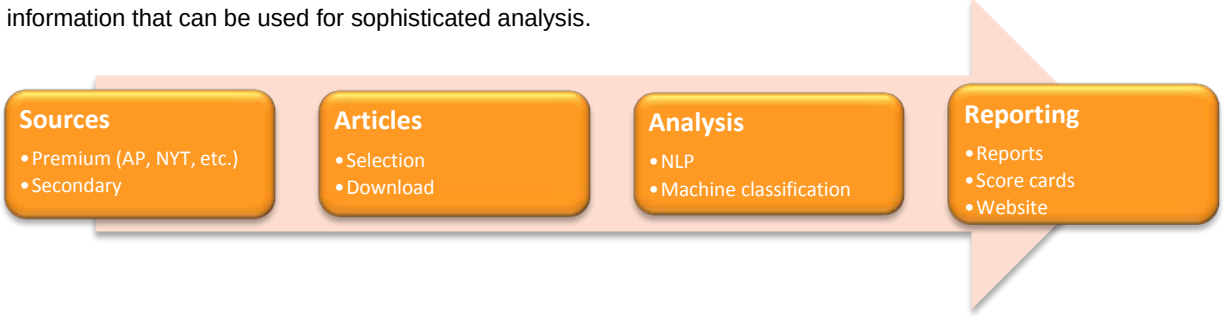
However, given the need to get the story out, and the diversity of news sources in America, how can the Department of Health and Human Services (HHS) know how well AP is following its own guidelines, or whether other news agencies and even blogs are following suit? Manual news searches and text analysis methods do not scale, and simple automated approaches cannot detect the meaning or subtle variations in phrases. NORC at the University of Chicago and the Entertainment Industry Council (EIC) provide a soluction that gives HHS the scoop on how the mentally ill are being represented in the media and blogosphere.

READING THE NEWS

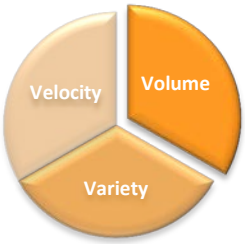
It's easy to read the news these days. Most newspapers are online, and search engines like Google aggregate news content, making it more accessible than ever before. However, manually searching for news articles for the many ways mental illness is represented is time consuming and error prone. It's difficult to acquire much data this way, and you're likely to miss as much as you find.

For a number of years now, organizations have tried to automate the process by selecting a few news websites and *scraping* (extracting and downloading) their contents. This may work for some free sites, but "premium" news sites now hide their content behind *pay walls*, and others forbid the use of automated tools altogether. Additionally, as with manual collection, it's difficult to collect enough valuable articles in order to support rich content analysis. Finally, if you do succeed in getting enough data to support research, then you still need to provision enough server hardware to store and analyze the data.

The NORC/EIC solution succeeds where traditional methods fail by utilizing innovative new news and social media APIs (application programming interfaces) to identify the best news sources and most valuable articles, and to "read" the news as it's published from the AP and many other influential news outlets, converting and persisting large volumes of it and converting the raw article text into natural language information that can be used for sophisticated analysis.



In addition to using innovative new social media and news APIs to identify and capture the most relevant news articles, this platform could be extended to capture and analyze reader comments about the stories for use in gauging reaction to stories and destigmatization efforts.

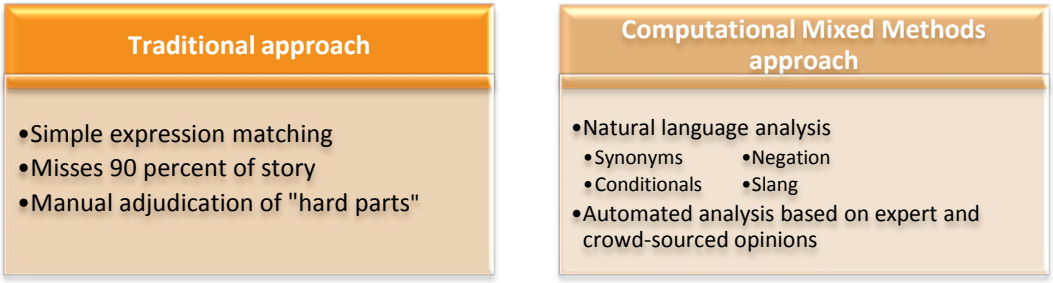


Big data requires big storage, and media analysis is no exception. NORC manages the Data Enclave, a state-of-the-art Managed Application Hosting Center (MAHC), providing the storage and bandwidth necessary to process gigabytes or even terabytes of news data on open-source, distributed file systems such as HDFS (Hadoop), and to host a web application where HHS employees can search and analyze the news. Additionally, the team has extensive experience working on leading cloud platforms, such as Amazon Elastic Compute Cloud (EC2) and Electronic Elastic MapReduce (EMR).

ANALYZING THE STORY WITH BIG DATA

Traditional text analysis methods involve using basic pattern matching to identify, extract, and encode data. For news analysis, this may mean ingesting a collection of terms with positive, neutral, or negative connotations related to the subject matter and reporting the polarity of the articles via basic statistical methods.

This is a good first start, but often misses much of the story. Text matching alone can't account for the complex grammar and inflection employed in news articles.

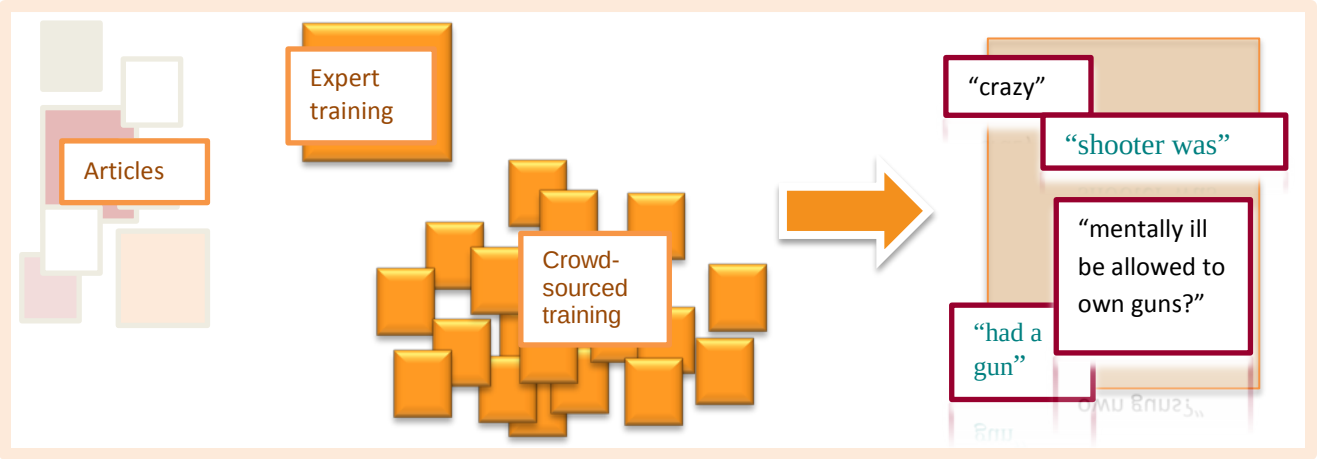


Enter NLP. Natural Language Processing permits rich analysis of text far beyond pattern matching. NLP uses sophisticated computational techniques and machine learning algorithms to account for negation, inflection, and variations in terminology to uncover the *meaning* of textual content. Additionally, new text processing tools allow for text analytics at scale, enabling the use of stemming, tokenization, full-text search, and other NLP techniques on gigabytes or terabytes of data in parallel using technologies such as Map Reduce. In particular, machine classifiers can be trained to automatically determine whether particular text documents belong to specific expert-defined categories—in this case, whether new articles conform to the new AP standards on mental health stigmatization. Algorithms such as these will allow us to monitor where and how mental health is being stigmatized in the media and track how such stigmatization changes over time.

The team has several years of experience in analyzing text to find its true meaning using leading tools, including the dominant open-source text analytics technologies such as Hadoop/Map-Reduce, Natural Language Tool Kit (NLTK), and Python.

EXPERT COMMENTARY

In order to properly train our data mining and natural language processing algorithms to interpret incoming news article data, we will use a diverse sample of extracted news stories and manually code them in terms of relevance to mental health issues and whether or not they conform to AP style guidelines. Using a combination of content expert coding as well as automated crowd-sourcing technologies, we will be able to develop high-quality classifiers that can effectively examine the distribution of stigmatizing language in the media and track how different forms of stigmatization change over time.



Combining these classifiers with other algorithmic techniques such as topic modeling will also allow us to break news articles into linguistic themes and measure the specific ways in which mental health is being stigmatized within the content produced by particular news organizations.

REPORTING THE NEWS

Text analytics provides the answers, but how do you ask the questions? To understand what is being reported, and how it's being reported, you need the ability to find what you're looking for and even have the system help you along the way with innovative search and analytics features.

Our team will produce a web application for searching, analyzing, and reporting on the data using leading open-source tools. The website will integrate with HHS security and authentication and authorization mechanisms, featuring an intuitive dashboard providing users with features such as full-text search, faceted navigation, saved searches, maps, alerts, rich reports, and other tools necessary to understand the effectiveness and reach of media-based destigmatization efforts.

ABOUT THE TEAM

NORC at the University of Chicago is an independent research organization that collaborates with government agencies, foundations, educational institutions, nonprofit organizations, and businesses to provide data and analysis that support informed decision making in key areas including health, education, economics, crime, justice, and energy.

Entertainment Industry Council (EIC) was founded to encourage accurate depictions of health and social issues in the media, acting as a bridge between the entertainment industry and public policy decision-makers.