Veracity Scoring of Social Media Content

Mark Wolff and Michael Wallis

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina, USA
E-mail : *mark.wolff@sas.com*

Abstract
Massive amounts of unsolicited, self-reported data related to consumer health, nutrition and preference are being generated everyday by individuals using a wide range of internet and social media channels. This "embarrassment of riches" shows no sign of moderation as the Internet of Things (IoT), including connected and wearable devices, will soon add to the present burden of data volume and heterogeneity. With the ever increasing ratio of computational power to cost and rapid developments in the automated analysis of unstructured data, text analytics and visualization, there exists an unprecedented opportunity to apply innovative approaches to gain valuable insights from the wealth of previously analytically unavailable data. A unique challenge arises related to the applicability and utility of data collected from web-based media sources. Adoption of these data as a resource is hampered by concerns related to their accuracy and reliability. We believe that the ability to collect these data en masse and evaluate the content for "veracity" would be of significant benefit and enrich understanding of the consumer with regard to the development of products focused on health and nutrition.

Keywords: Social Media, Text Analytics, Unstructured Data, Data Veracity, Data Quality.

1. Introduction

The internet and its reach beyond the desktop via mobile devices, has become a dominant and ever present source of information for many around the world. The development and adoption of Web 2.0 technology and social media applications have further accelerated this trend. Vast amounts of data that are shared and accessed on the internet are unstructured. These data exist as documents, wikis, blogs, comments, tweets, Facebook status updates, crowdsourcing sites, etc. The wide spread popularity of photo sharing sites and the ability to tag images and video has further added to the available pool of unstructured data.

As more and more individuals, organizations and institutions rely on the internet for information to support important decision making, the question of data integrity and veracity becomes ever more critical. Whether a consumer making an important personal health care decision or company developing and marketing a novel health related product or the government or other regulatory bodies, monitoring the internet has become part of our daily experience and has become a routine and important component of decision making.

A particular area if interest and concern is the applicability and utility of social media data to help better understand the risk to benefit of products. Individuals routinely contribute and consume solicited and unsolicited information about a variety of health related issues using a wide range of internet and social media channels. Such data offer a potentially valuable resource in the context of monitoring risk, benefit and preference for the industry and for regulatory agencies. Adoption of these data as a resource has been hampered by concerns related to the accuracy and reliability of data and a lack of guidance for the industry especially as related to issues of legal liability. We believe that the ability to evaluate and score information for veracity of social media data would provide a benefit to public health and safety, to the pharmaceutical industry and to the regulators. Each would benefit from the ability increase confidence in data collected from social media.

We propose a method for applying a "veracity score" to data that have been collected and prepared for analysis from a variety of internet sources, including social media. The "veracity score" would be in context of the source data and would reflect a measure of confidence associated with the specific data collected. Unstructured data would be processed through a Natural Language Processing/Text Analytic engine. The analysis entails the application of both empirical and theoretical methods to understand how specific relationships between the data, its source and other factors can be used to evaluate "veracity". Additionally, a process of logical analytical gates, based on rules, outlier detection, and pattern analysis/ matching can be applied to further verify and score the collected data in preparation for safety related analysis.

For example, a forum entry that contained self-reported information about an individual's experience with a specific medical device would be analyzed as to what relevant information are present. Those data would be processed with appropriate dictionaries and taxonomies and would be cross referenced with internal data as well as external databases. Summary reports would be generated identifying the source, relevant information such as author, data source, post/repost, date, etc. for data which have met established veracity thresholds. Thresholds and alerts can be set to identify specific cases for review thus creating an automated, learning-enabled system.

To demonstrate our approach of automated text analysis of publicly available Social Media and Unstructured Data sources from the web we have focused on an individual medical device, the Vagus Nerve Stimulator.

Our objectives were as follows;

1) Identify four websites where individuals are posting publicly available comments on their experience with the device of interest
2) Collect, processes and analyze comments
    a. Identify documents that;
        i. Refer specifically to the device of interest
        ii. Comments that are part of a thread related to the device of interest
        iii. Contain terms identified as known "Adverse Events" (AEs)
            1. AE terms consistent with the product label
            2. Specific AE terms not on the product label
            3. Terms that identify other drugs, substances or devices
    b. Perform document author "Veracity/Integrity analysis
3) Identify relevant scientific literature PubMed abstracts
4) Identify relevant reports as collected on the MAUDE database

The collected and processed information will be analyzed and reported as frequencies of terms over time. Our goal was to establish a historical baseline for frequencies of specific AE terms and related. Such a base line would establish a reference point by which disproportional increases or decreases in frequency could serve as an "alert" that may point investigators toward a specific area for further analysis. Additionally, the system would be able of identifying terms as specified by the investigator for monitoring and alerting and searching the collected documents.

The information collected from web-based sources can be cross referenced by similar methods as applied to abstracts in MedLine and on unstructured content available in the MAUDE Data base.

Summary of analytic approach to veracity & data integrity;

1. *Rules*
    a. Rules and thresholds based on known behaviors/patterns
    b. Biological/Clinical Plausibility
    c. Product label
    d. IP Address/URL
    e. Author ID

2. *Anomaly Detection*
    a. Unknown Patterns and Behaviors
    b. Algorithms used to detect unusual patterns
    c. Multivariate outlier/inlier detection
    d. Constant findings
    e. Clustering/association analysis
    f. Distribution analysis

3. *Predictive Modeling*
    a. Complex Patterns
    b. Identify patterns which describe inaccurate information
    c. Apply unsupervised/supervised learning techniques
    d. Like patterns of comments and content
    e. Author verification
    f. Higher level concept disambiguation

4. *Network Analytics*
    a. Clustering/Associative Linking
    b. Discovery through automated link analysis
    c. Understand complex multivariate relationships over time
    d. Link authors to suspect content

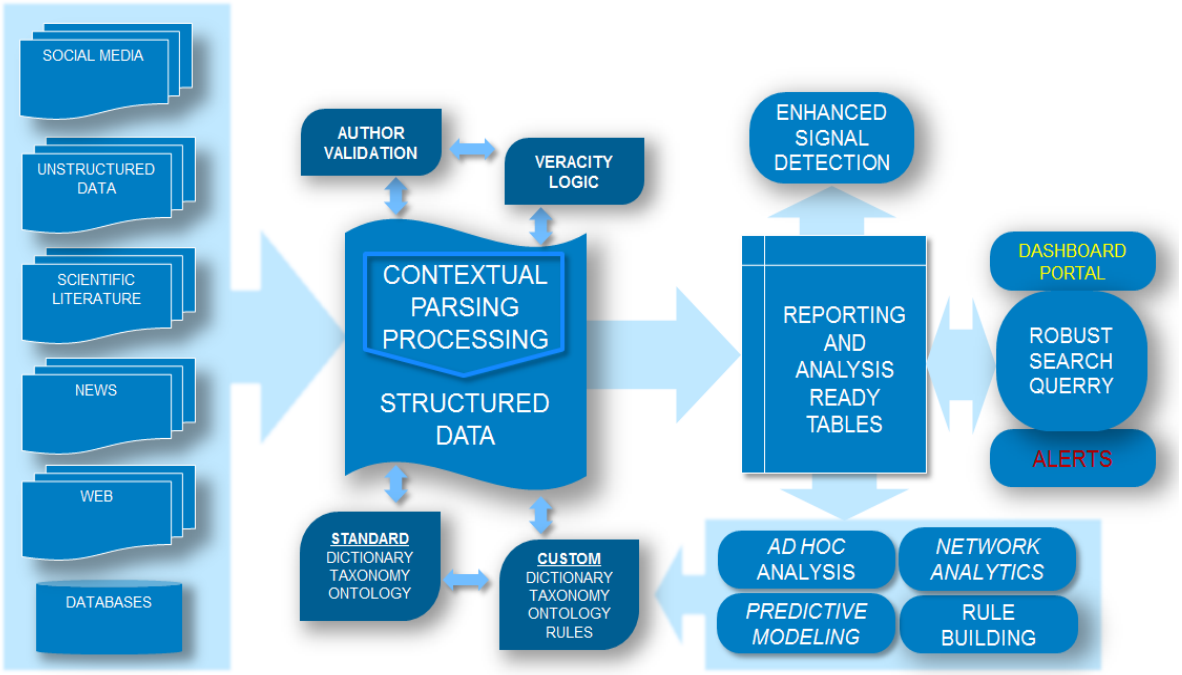

Figure 1. Project Schematic.