Reviewing Asthma-related Grey Literature and Personal Opinions on Twitter using LDA and CTM Clustering

Gondy Leroy, PhD, Joe Koolippurackal, Shikha Swami, and Philip Harber, MD MPH

University of Arizona, Tucson, Arizona, USA

Abstract
Twitter is a big data source useful for sampling opinions and information in the community. In this exploratory study, we compared the usefulness of information shared about asthma. We collected tweets using asthma and #asthma and compared tweets with and without URLs. To automatically create topic overviews that can be efficiently reviewed, we used LDA and CTM clustering. We evaluated using 2, 5 or 10 clusters.

Introduction
Social media play an increasingly important role in medicine. We focused on asthma, which affects 8% of the US adult population(1), and analyzed Tweets to assess: (a) what proportion is unique; (b) differences between tweets from individuals vs grey literature (i.e., those referring to professional information); (c) whether searching for tweets with or without a hashtag matters.

Methods and Results
In fall 2016, we collected 178,796 tweets for asthma and #asthma using the Twitter Search API. Based on prior work(2), we made the broad assumption that tweets with a URL refer mostly to grey literature tweets (by professionals), while tweets without URL are more often comments by individuals. We removed duplicate tweets and stopwords, then stemmed the terms. To review tweet content, we compared Latent Dirichlet Allocation (LDA) and Correlated Topics Model (CTM) clustering for 2, 5 and 10 cluster.

Overall, 70% of the tweets were unique (Table 1). Tweets with a URL are 64-77% unique. Tweets without URLs were more unique and almost completely unique for asthma. This is not unexpected since this query is broad (appearance of word asthma) and personal comments can be expected to be unique. To evaluate the content of the entire set in an efficient manner, we clustered the tweets and evaluated the clusters. All cluster sets were shown in random order and evaluated (blind) by a domain expert, who scored each term on a 4-point scale of relevance to asthma (an asymmetric scale was used to penalize nonsense terms). Figure 1 and 2 show the average normalized scores. Scores for all terms were combined per condition. LDA clustering resulted in more relevant clusters and more consistent quality. Tweets for #asthma (with hashtag) by individuals contained the most relevant terms. Terms were least relevant for the asthma query without URL (personal comments, broad query). Implication: For estimating community members’ opinions & information, relevance is improved by use of LDA clustering of tweets without URL references.

References