Social Network Analytics Data Mining Applied To Twitter Posts

[1,]Prasad V. [2,] Dr. T. Nalini

[1,]*SAP Consultant/SAP Basis support Kavaian System PVT LTD*
[2,]*Professor, Computer Department Bharath University*

***Abstract**: The social media has grown significantly in recent years. With the growth in its use, there has also been a substantial growth in the amount of information generated by users of social media. Insurers are making significant investments in social media, but many are not systematically analyzing the valuable information that is resulting from their investments. This paper discusses the application of correlation, clustering, and association analyses to social media. This is demonstrated by analyzing insurance Twitter posts. The results of these analyses help identify keywords and concepts in the social media data, and can facilitate the application of this information by insurers. As insurers analyze this information and apply the results of the analysis in relevant areas, they will be able to proactively address potential market and customer issues more effectively.*

I. INTRODUCTION

Regardless of where you look, you can see an explosion in the use of social media.

Online communities have developed that focus on both personal and professional lives. Groups have been formed that focus on every potential area of interest, including food, sports, music, parenting, scrapbooking, and actuarial issues. It is estimated that there are over 900 social media sites on the internet. Some of the more popular platforms are Facebook, Twitter, LinkedIn, Google Plus, and YouTube.

- People spend over 500 billion minutes per month on Facebook.
- There are 200 million registered Twitter accounts.
- There are more than 70 million users of LinkedIn worldwide.
- YouTube receives more than 2 billion viewers per day.
- Seventy-seven percent of internet users read blogs.

The majority of the population is using social media in some form or another.

Given the substantial increase in the use of social media, there is a significant amount of information that is, being generated. As seen in the same sources referenced above, the volume of this content is staggering:

1.1Research Context

The context of this research will focus on data and text analytics. Since much of the data from the social media sites will be text-based data, the process of preparing and analyzing the data will focus on principles of preparing text data for analysis. The author was unable to find anything in CAS literature that focuses specifically on the analysis of social media. However, a good discussion of the principles of text mining in CAS literature is in a paper written by Louise Francis entitled “Taming Text: An Introduction to Text Mining.” [5] Building on these concepts, there are some unique considerations when analyzing text data from social media sites which will be discussed in this paper.

1.2.Objective

The purpose of this paper is to describe, through the use of a specific example, how data miningand text analytics can be applied to social media to identify key themes in the data. Specifically, this paper will describe the analysis of Twitter posts related to the keyword Allstate.Allstate was chosen purely based on the public availability of historical Twitter data. While this example helps to make some of the points and concepts clearer, the purpose of this paper is not to provide a detailed analysis of Twitter activity related to Allstate, but to demonstrate how analytics can specifically be applied to social media information related to a property and casualty insurance company.

II. TWITTER BACKGROUND AND DATA DESCRIPTION

Twitter is a social networking site that allows users to send and read short messages of a maximum of 140 characters. Twitter was created in March 2006 and was officially launched in July 2006. The growth of Twitter has been phenomenal, currently having reached over 200 million users and handling over 200 million tweets per day. Users sign up for an account on Twitter, and once they have an account they can begin to “tweet,” which is the terminology for sending a message. Users can subscribe to other user’s tweets, a process known as “following.” These subscribers are known as “followers.” By default, tweets that a user sends are visible to everyone; however, users can also choose to send tweets specifically to their followers that will not be visible to the public. Users on Twitter are identified by a user name, and this user name is proceeded by the "@" symbol. When a user identifies another user in their tweet by their user name, it will be visible to the public, and the user that is referenced will be notified by Twitter that they have been "mentioned." If a user sees a tweet that is interesting and wants to pass the information along, they can "retweet" the post, which is similar to forwarding an email message to a new set of users, in this case their followers. Retweets will generally be identified with an "RT" that is embedded in the message. Lastly, messages can be grouped by topic or type by the use of hashtags (#). A hashtag preceding the topic will allow Twitter users to find tweets related to a particular topic when performing a search.

The data that was captured from twapperkeeper.com includes the following information:

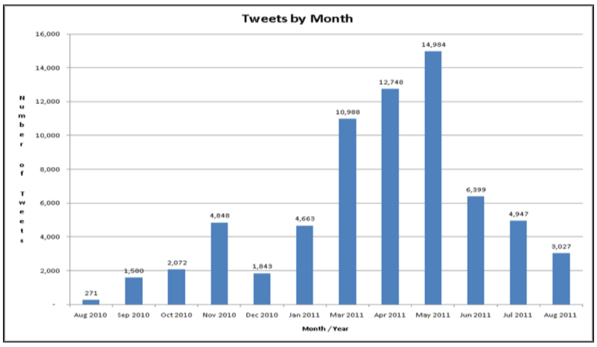
The data that was captured from twapperkeeper.com includes the following information:

- User : the username that sent the tweet
- Tweet : the content of the tweet
- Timestamp : the date and time the tweet was sent (GMT)
- Tweet ID : Twitter identification number of the tweet
- Geo : latitude and longitude of the user

It must also be remembered that this data was captured based on the use of the hashtag Allstate. Therefore, it will not capture every tweet that uses the word Allstate, but rather those tweets where the user specifically identified Allstate as a keyword. Also, twapperkeeper.com makes no guarantees that they capture all tweets that meet the archive criteria, so there could potentially be tweets with #allstate that were not captured. While this may introduce a bias, the concepts for analyzing the tweets are still valid.

III. GENERAL DESCRIPTIVE STATISTICS

There are a total of 68,370 tweets that were used as part of this analysis. The tweets used began on August 1, 2010 and ended on August 12, 2011. The number of tweets by month is shown below.



As can be seen in the figure above, the number of tweets captured per month varied between 1,500 and just under 5,000 through January 2011, at which point the number of tweets increased to 10,000 – 15,000 per month for March through May 2011. June and July settled back to pre-March, 2011 levels. August 2011 only represents 12 days of tweets, so it was not a complete month as of the time this paper was written.