Build social media datamarts using SPSS text mining tools

Kimberly Chulis  
October 02, 2012

The rise of social media has changed the way big brands do business. Prior to the widespread adoption of social channels, consumer behavior was based more on independent information gathering and preference formation. Customers are increasingly online, engaging in conversations, asking for advice, performing price and quality comparisons, and influencing others. These individual-level behaviors embedded in raw social media data represent consumer preference, purchase history, significant life events, mood, personality, and other attributes that can be derived through text mining and stored in a social media datamart.

Introduction

The precursor to the social networks we know today emerged in the late 1960s, when bulletin boards were one of the first interactive message-sharing platforms. It wasn't until more recently—in the 1990s, when craigslist and AOL arrived on the scene—that the social revolution gained rapid ground. Social networks took off in the 2000s, with Friendster, LinkedIn, MySpace, Flickr, Vimeo, YouTube, and then Facebook in 2004 and Twitter in 2006, and most recently Google+ and Pinterest.

The digital trends that accompany the widespread adoption of social media have direct implications for brands as they develop a fluid digital strategy for an environment that is characterized by moving parts. The social stream is effectively lengthening the relationship between brands and customers. Prior to e-commerce and social media, consumers did some research about products and made a purchase that was distinct, and the relationship ended until it was time for a subsequent purchase. Word of mouth was limited to a consumers' physical social network. Now, customer opinion is amplified through social networks, with a potential reach across the entire consumer audience.

Brands know that today's consumers are actively gathering pre-purchase information, where they review other favorable or unfavorable opinions and are better able to perform rapid price comparisons with a few taps on a mobile device. They also know that their consumers are far more sensitive to the influence of others in their social network, which is leading to the development of a new type of influencer loyalty program aimed at incentivizing and rewarding
individuals who wield powerful brand influence. Customers are becoming the new brand champions in such a way that aligning brand personality and brand identity has never been more critical for brand survival.

How, then, are brands managing this influx of digital interaction information? Technology has been racing to catch up with the rise of the social consumer. The social networks themselves have provided site-specific traffic and stats tools, such as Facebook Insights, YouTube Insights, and social media management suites like HootSuite, and influencer measurement portals like Klout provide third-party options for brand engagement tracking metrics. A variety of commercial social listening tools, such as Radian6, SM2, Viralheat, and Sysomos, provide reporting, text analytics, engagement, sentiment analysis, visitor information, and engagement workflow. These tools are improving in scope and usefulness, but many of them are still in an early stage of evolution. Sentiment analysis, for example, is still far from accurate, and social data provided through services such as the Twitter firehose, and by partner companies such as Gnip and DataSift, are still prohibitively expensive and limited in the nature of data available. Therefore, there is a strong argument for augmenting these commercial tools with in-house text mining and the construction of a proprietary social media datamart. Social media datamarts store consumer-level information derived from social media interaction and all of the associated digital information around location, device, mobile behavior, mobile payment, platform, and speed related to the comment data.

Text mining and semantic methods

Given that social media generate a wealth of consumer data, how can brands turn raw social media comment data from Twitter, Facebook, blogs, and forums into actionable business insights? The answer lies in the application of text-mining and semantic technology to these new sources of unstructured data.

Text mining refers to the techniques used in the extraction of information from different written sources. Why is this so important? It is widely estimated that 80 percent of all business-relevant information resides in unstructured and semi-structured text data. In other words, without the application to text analytics to unearth the wealth of data represented in that 80 percent, all of the embedded business information and consumer behavior data goes to waste. The term text mining, often referred to as text analytics, has many practical purposes, such as spam filtering, extracting information from suggestions and recommendations on e-commerce sites, social listening and opinion mining from blogs and review sites, enhancing customer service and email support, automated processing of business documents, e-discovery in the legal field, measuring consumer preference, claims analysis and fraud detection, and cybercrime and national security applications.

Text mining is similar to data mining in that it is aimed at identifying interesting patterns in data. Although manual (and highly labor-intensive) text mining emerged in the 1980s. The field of text mining has become important in recent years for refining search engine result algorithms and sifting through data sources to essentially discover unknown information. Techniques such as machine learning, statistics, computational linguistics, and data mining are all employed in the process. The goal of knowledge discovery from text, for example, is to detect underlying semantic relationships from text as well as content and implied context with Natural Language Processing (NLP). The processes are aimed at using NLP to replicate, and then scale the same kind of
linguistic distinction, pattern recognition, and resulting comprehension that occurs when human beings read and process text.

Various methods exist in the field of text mining. The following introduces a list of common and sequential steps involved in text mining.

The first step in any text-mining effort is to identify the text-based sources to be analyzed and gather this material through information retrieval or selecting the corpus that comprises the set of textual files and content of interest. Extensive NLP is deployed that invokes "part of speech tagging" and text sequencing to parse for syntax (that is, tokenizing text) and applying Named Entity Recognition (that is, identifying the mention of brands, people's names, places, common abbreviations, and so on). An iterative Filter Stopwords step involves the removal of stopwords to refine desired topical content. Pattern Identified Entities recognizes email addresses and phone numbers, and Coreference identifies noun phrases and related objects in text, followed by Relationship, Fact and Event Extraction. N-Grams are often generated that create terms as a series of consecutive words. Finally, sentiment analysis, an approach used widely by social media listening and categorization tools today, is performed to extract attitudinal information toward the object or topic. Often times, various mapping and plotting functions provide visualization for further accuracy validation.

Text mining tools

There are several commercial and open source options for text-mining software and applications. IBM offers a wide and robust variety of text mining solutions. A powerful offering that leverages the big data capabilities of IBM® InfoSphere® BigInsights™, provides an add-on text analytics module that runs text analytics extraction from the InfoSphere BigInsights cluster. The IBM SPSS® offerings range in scale and scope. One tool that works well for searching a document and assigning it to a topic and subject is IBM SPSS Modeler, which provides a graphical interface to perform generic text document classification and analysis. Another product, IBM SPSS Text Analytics for Surveys, uses NLP and is useful for analyzing open-ended survey questions in a document. IBM SPSS Modeler Premium runs on the same engine as SPSS Text Analytics for Surveys but is highly scalable to handle an entire corpus of documents (PDF, web pages, blogs, emails, Twitter feeds, and more) in a sophisticated workbench that also facilitates integration between structured and unstructured data. A related custom source code node for Facebook extends the capabilities of SPSS Modeler Premium to read data directly from a Facebook wall and integrate it with a Twitter feed in SPSS Modeler to get a multi-social media channel perspective.

Of the open source text mining tools, RapidMiner and R appear to be two of the most popular. R has a wider user base; a programming language in which source code is required, it has a large selection of algorithms. However, scalability is an issue with R so it's not ideal for large datasets without workarounds. RapidMiner has a smaller user base, but it doesn't require source code and has a powerful user interface (UI). It's also highly scalable and can handle clusters and in-database programming. IBM offers a Jaql R module that integrates the R project in queries, which in turn allows MapReduce jobs to run R computations in parallel.
Social media datamarts and big data

NoSQL and SQL options
A quick word on NoSQL and Structured Query Language (SQL) options and the technology stack selection process. When data sources become unwieldy, as they frequently do with social media data, a combination of commercial NoSQL options (such as IBM InfoSphere BigInsights) that integrate efficiently with Hadoop and other capability-extending open source tools is necessary. Graph databases and key-value and document stores are available, and the optimal selection is based on primary use case. Companies interested in text mining and analytics typically select Hadoop and integrate with other open source tools such as Apache Mahout, a machine learning engine that provides classification, clustering, and collaborative filtering. Storm's tuples and steams manage real-time analytics that navigate Hadoop's high latency.

Unique challenges exist when setting out to apply text mining to social media data. The data that social networking sites, blogs, and forums generate falls in the category of what is commonly referred to as big data. The data is unstructured and semi-structured, petabytes are generated around larger brands on a daily basis, and traditional relational databases cannot efficiently scale to support real-time analytics based on the data. Big data and NoSQL database solutions are therefore required.

Social media data, if not collected and adequately stored at regular intervals, is essentially perishable. Most open source social listening tools only store a few days' worth of social media comment history. Twitter only recently announced that an entire history of data will be available, but it will be limited to comments posted specifically by the account holder. This data is available from some of the larger social data providers mentioned above, such as Gnip and DataSift, and through volume and call-based application programming interfaces (APIs) through other tools. However, where it is available (for Twitter), it is prohibitively expensive for all but the largest brands.

Each social media site handles this issue differently. It is possible to use search requests and have JavaScript Object Notation (JSON) format responses that contain unparsed data for immediate inclusion in either a MySQL or NoSQL database, depending on volume and the nature of the data.

Business use cases for text mining
Brands have different objectives with text mining exercises:

- A company like Sears, in Example 1, may be interested in tracking customer sentiment through social media comments and Facebook page fan interactions directly following the launch of a new product line. In this way, it is possible to understand basic sentiment around pictures, products and the conversation clusters happening around the product launch. This real-time feedback allows for rapid message updates and removal of unpopular content, and Facebook fans become a real-time focus group, providing immediate feedback on product features.
- A company like JACT Media is in the business of building relationships between brands and video game players. It has an in-game overlay that allows gamers to play their regular games while displaying a variety of targeted, scheduled content to players. Gamers earn
JACT virtual currency, and these JACT BUX can be redeemed for rewards, including virtual and downloadable goods. Players interact with JACT on the Facebook page or Twitter, and mention JACT BUX frequently on game forums. This raw comment data can be harvested from the various sources, and individual-level comments and preferences can be stored. For instance, if a player is excited about a particular video game or tweets about his or her reward, in-game targeting based on specific game and reward type is more likely to drive increased loyalty than random offers.

- Supermarkets are able to use social media data to identify more valuable shoppers, impressions of customer service, store atmosphere, product preference, packaging preference, and pricing. Merging this type of information with location data that either Twitter or mobile devices provide, supermarkets can custom tailor the shopping experience from a localization perspective. This has implications for inventory, pricing, advertising, individual digital and direct mail coupon offers, and more.

**Example 1: Social media data and text mining in SPSS Modeler**

This first example shows a use case for SPSS Modeler. In this scenario, a new product line is launched, and the company is interested in tracking consumer response in social media data. A Facebook node, which was developed using the IBM SPSS Modeler Component-Level Extension Framework (CLEF), is used to track this new Kardashian product line on the Sears Facebook page, shown in [Figure 1](#).

**Figure 1. Retailer launches a new product line on Facebook**

![Facebook Page](image)

The first step in tracking and analyzing comment data involves the user specifying a user name and number of pages and threads for review in the Facebook node, shown in [Figure 2](#).
The comment data is then extracted from the Sears Facebook page and made available for use in SPSS Modeler, as shown in Figure 3.

Figure 3. Raw comment data can be viewed directly via the SPSS Modeler Facebook node

The next step involves adding filters and performing concept extraction, resulting in a visualization that depicts the content categories around the brand. The user-friendly graphical UI guides the user through the process, and no APIs are needed to extract the social data from Twitter or Facebook. What results is an easy-to-comprehend concept map and sensitivity to concept clusters represented by the thickness of the connecting line, as shown in Figure 4.
Figure 4. Concept Map provides visualization of strength-of-concept categories to brand.

Example 2: Supermarket product preference example using extraction and stopwords in SPSS Statistics Base

The following social media datamart assembly process describes a simple manual text mining process. In this example, we are interested in using text mining through SPSS Statistics Base to derive and store individual product preference from social media data. The example includes a stepwise guide to extracting supermarket brand data from Twitter and Facebook. The process architecture is represented in Figure 5.
The first step is to identify the brands of interest. A routine is set up to collect brand-related mentions through an API process. This is done with search requests such as those depicted in Figure 6, and results are returned in JSON format. A JSON library parses the data, and each record is split up into multiple fields that contain such information as user ID, data, and unprocessed textual message comment. This data is then stored in a database and made available for text mining.
Figure 6. Sample API to access raw Twitter and Facebook comment data

Social Media Data source: Facebook/Twitter

1. Search request to Facebook and Twitter
   - http://search.twitter.com/search.json?q=kroger
   - https://graph.facebook.com/search?q=kroger&type=post

2. The app receive the results in a JSON format

   ```json
   {
     "id": "100000741841149_448892445145480",
     "from": {
       "name": "Frank Kneal",
       "id": "100000741841149"
     },
     "message": "well no luck with kroger, still lookin....happy birthday little brother",
     "type": "status",
     "application": {
       "name": "Facebook for Android",
       "namespace": "fbandroid",
       "id": "350685531728"
     },
     "created_time": "2012-07-26T15:49:32+0000",
     "updated_time": "2012-07-26T15:49:32+0000"
   }
   ```

The objective of this simplified text mining exercise is to identify specific consumer product preferences and consumption patterns. This information is then stored in a social media datamart. For this specific example, suppose that you want to identify all of the customers who are consumers of the vegetable corn. Figure 7 shows the use of the Character Index function, which identifies all instances of the word *corn* in raw comment data.
What results requires further filtering, and stopwords are applied through various iterations to improve classification accuracy. By applying stopwords such as popcorn, candy corn, corndog, and corn syrup and limiting the instance to a four-character combination, a much more accurate identification of the corn product consumers results. These user names can then be flagged with a 'corn_consumer_flag'=1 in the database and selected for corn-specific offers and recipes in future marketing campaigns. (See Figure 8.)
When you have gone through an exhaustive list, you can then perform user ID aggregation and populate tables to capture product purchases, comments around packaging, and other variables that store individual-level consumer behaviors. In this example, raw social media data is stored in a NoSQL database, and the derived product preference flags are stored in a MySQL datamart, where user ID is a primary match key (see Figure 9).
Conclusion

Text mining is gaining in popularity as many businesses struggle to assess the potential return on investment of social media as a marketing and brand interaction channel. Companies are rushing to implement big data storage solutions to house unstructured data and integrate it with traditional transactional-type data. Social media comment and brand-related interaction data offers a wealth of insight into individual consumer preferences that can be used to design relevant product features, marketing in a way that will resonate with consumer desires and expectations. Storing this individual-level behavior and preference data in social media datamarts for the purpose of deeper brand experience customization will put information in the hands of a company that can be used to enrich the consumer-brand relationship and promote consumers to get engaged in self-management of their own brand experience.
Related topics

- **Ultimate timeline of social networks, 1960-2012**: Learn more about the history of social networks.
- **Social retail - finding, engaging and cultivating today's connected consumer**: Read more about how consumers gather pre-purchase information before buying.
- **2012 Compare best social media monitoring tools**: Companies use many social listening tools, such as Radian6, SM2, and Sysomos.
- **Text mining**: Find a good introduction hosted on Wikipedia.
- **Big data**: Learn more about its role in the enterprise.
- **developerWorks on Twitter**: Join today to follow developerWorks tweets.
- **R versus RapidMiner**: Analyticbridge provides a good comparison.
- **Jaql's R module**: Find more information on this module that enables you to integrate the R Project for Statistical Computing into your Jaql queries.
- **Evaluation software**: Find more trial software, including several SPSS products. Download a trial version, work with product in an online, sandbox environment, or access it in the cloud.
- **SPSS Text Analytics for Surveys** analyzes survey text to extract consumer insights.
- **Storm**: Check out this real-time computation system.
- **KDnuggets** provides a full list of commercial and open source text mining tools.

© Copyright IBM Corporation 2012
**Trademarks**
(www.ibm.com/developerworks/ibm/trademarks/)