Introduction

In this age, in this country, public sentiment is everything. With it, nothing can fail; against it, nothing can succeed. Whoever molds public sentiment goes deeper than he who enacts statutes, or pronounces judicial decisions. Abraham Lincoln, 1858

It is clear from President Lincoln’s famous quote that politicians realized the power of public sentiment a long time ago. Although keeping track of the public’s opinions and sentiments about topics and events of general interests has been in the domain of political and social scientists, marketers and businesses are typically interested in a consumer’s opinions and sentiments that specifically relate to their products, services, and so on. Monitoring what consumers are saying about a company’s brands and products and how they are expressing their opinions and sentiments to others has always been important to businesses. The difference now is the scale and scope of a consumer’s opinions and expressions and the toolset that is available to marketers compared to what existed just a few years ago. Until the last century, businesses typically used surveys and focus groups from time to time to gauge and track consumer sentiments. With the widespread adoption of the Internet, the proliferation of social media channels (such as Twitter, Facebook, and others), and the abundant opportunity for consumers to express their opinions and sentiments, monitoring sentiment continuously has become more critical. Conceptually, there is a difference between asking people questions about their feelings and sentiments and getting answers versus people expressing their opinions and sentiments freely without any prompting. Although companies still continue to capture a customer’s opinions and sentiments via regularly scheduled surveys and focus groups, those methods need to be supplemented by what is freely expressed by consumers on the Internet, particularly in social media.

“Conventional marketing wisdom long held that a dissatisfied customer tells ten people; but in the age of new social media, he or she has the tools to tell millions,” says Paul Gillin, author of The New Influencers: A Marketer’s Guide to the New Social Media. Therefore, it becomes imperative for companies to actively monitor conversations on the web among consumers about a company’s brands, products, or services, and to glean information about what positive and negative sentiments are being expressed. This is a challenging task for many reasons, such as too many sites on the Internet, too many formats in which opinions and sentiments are stored and displayed on the web, idiosyncrasies in expressed opinions that are both time and domain dependent, and so on. In this chapter, sentiment analysis is used as an umbrella term that covers the analysis of people’s expressed opinions, expressions, emotions, and sentiments about brands, products, services, events, topics, and their attributes (Bing Liu 2012). Sentiment analysis (sometimes referred to as opinion mining in academic literature) is a field of study that originated in computer science, but has gained traction in business and marketing literature. In the past few years, many applications of sentiment analysis have been published using a variety of textual data and business contexts, such as tweets about movies (Castellanos et al. 2011), reviews about movies (Joshi et al. 2010), tweets about events related to retailers (Duraipadayal et al. 2012; Grover et al. 2013), and reviews about products and services (Liu et al. 2013; Nangiajan et al. 2015; Pantangi et al. 2015; Sarkar et al. 2013).
Basics of Sentiment Analysis

The basic task involved in sentiment analysis is identifying and quantifying the polarity or valence of sentiments (such as positive, negative, neutral, or mixed) expressed typically in written opinions, expressions, reviews, comments, and so on. Therefore, it involves many of the text analytics steps such as tokenization, sentence identification, part-of-speech tagging, and so on, that have been discussed in previous chapters. But, sentiment analysis has to go beyond the basic steps in text analytics. For example, often tweets, reviews, and postings have different types of sentences (declarative, imperative, and interrogative) that must be identified because they might not be appropriate for sentiment analysis. For example, a declarative sentence (such as “It is an amazing TV.”) often states the views of the author and is appropriate for sentiment analysis. An imperative sentence (such as “Do not buy this TV.”) can be used to infer sentiments about the product. However, an interrogative sentence (such as “Which is the best TV?”) might not be a good candidate for sentiment analysis. Depending on the context, sentences can be non-comparative (where opinion is restricted to one thing) or comparative (where multiple things might be compared). Comparative sentences are more difficult to handle than non-comparative sentences. Just as in text analytics, sentiment analysis requires the analyst to spend considerable time preprocessing text and making a large number of subjective judgments to get meaningful results.

Sentiment analysis starts with determining whether a text contains an opinion (sentiment). If it does contain sentiment, at what granularity level does the sentiment exist? Consider the following example of a review by a customer for a TV:

The TV is wonderful. Great size, great picture, easy interface. It makes a cute little song when you boot it up and when you shut it off. I just want to point out that the 43” does not in fact play videos from the USB. This is really annoying because that was one of the major perks I wanted from a new TV. Looking at the product description now, I realize that the feature list applies to the X756 series as a whole, and that each model’s capabilities are listed below. Kind of a dumb oversight on my part, but it’s equally stupid to put a description that does not apply the listing for a very specific model.

Many questions come to mind as you read the review. There are some statements that are clearly subjective and contain an opinion, and other statements are just facts. Considering the opinion statements, is the overall sentiment (in the entire review) about the TV expressed by the customer positive, negative, or mixed? What about the sentiment for attributes of the TV, such as size, picture, and interface? What about the video-playing ability of the TV from the USB? A clear understanding of the level of analysis and the terms used to specify these levels is needed. Liu (2012) defined three granularity levels for any sentiment analysis problem as follows:

**Document level:** At this level, the task is to figure out whether the entire document can be classified as positive or negative. This is possible only if the document involves a single entity (such as the TV in the previous example). If a document involves multiple entities, then it might not be possible to arrive at a document-level sentiment classification.

**Sentence level:** At this level, the task is to classify each sentence in a document as positive, negative, or mixed sentiment sentence. In the previous example, the first sentence, “The TV is wonderful.” expresses positive sentiment. The third sentence, “I just want to point out that the 43” does not in fact play videos from the USB.” expresses negative sentiment. Depending on the nature of the text, some sentences might not express any sentiment and therefore cannot be classified.

**Entity (or Object) and Attribute (or Aspect or Feature) level:** An entity is typically the target of the opinion. In the previous example, the target of the review is clearly the TV, which is the entity for the entire review. However, in many sentences, the sentiments reflect the reviewer’s opinions about attributes (or aspects or features) of the entity. For example, in the second sentence, (“Great size, great picture, easy interface.”), positive sentiment is being expressed for three specific attributes (size, picture, and interface) of the entity, the TV.

Often a hierarchical taxonomy is used to represent an entity and its various attributes. Sentiments can then be applied to each attribute. In this example, the TV is the root entity (object). Each root attribute (size, picture, and interface) is a component or subcomponent of the root entity or object. You can express an opinion about the root node or the object (TV) or any components or subcomponents. In SAS Sentiment Analysis Studio, the term “feature” is used instead of attribute and the term “product” is used instead of entity. These terms are used hierarchically in the sense that features are nested in a product. However, in SAS Sentiment Analysis Studio, it is possible to define intermediate entities that frequently occur in the documents so that they can be referenced and used again in the same analysis.

Challenges in Conducting Sentiment Analysis

Sentiment analysis starts with text data. As a result, all of the typical natural language processing (NLP) problems associated with analyzing text data, such as correctly identifying part-of-speech tags, disambiguating terms and lexicons, correcting spelling errors, and so on, can plague sentiment analysis. The commonly used lexicon-based sentiment analysis approach depends on correctly identifying opinion words that express positive or negative sentiments in a sentence. There are general opinion words whose polarity is always the same, independent of the context. An example is the word “beautiful,” which expresses a positive sentiment. These are usually easy to handle. But, there are also context-dependent lexicons in which the polarity of the word depends on the domain or context. For example, the word “small” can be positive or negative depending on the context. The sentence, “The size seems small.” can be positive for a USB flash drive with 1 TB capacity. But, the same sentence can be interpreted as negative if the context is an LED big-screen TV. In addition to opinion words, there are idiom lexicons—typical expressions such as “costs an arm and a leg” that embody sentiments. In general, the difficulty with correctly identifying sentiments increases as you move from general to context-dependent to idiom lexicons in texts.

Other challenges in conducting sentiment analysis arise from the nature of the text that is being analyzed. For example, tweets are short, and they are typically focused on one topic only. In that sense, they are easier to analyze. But, tweets often contain a lot of special meaning characters, such as RT (retweets), hashtags (#), emoticons (such as smiley faces), that need to be handled carefully. Customer reviews are typically one entity or object. Therefore, there is less ambiguity in the entity detection task when analyzing reviews. Analysis of discussions, free-flow comments, and blog postings is often the hardest because they typically cover multiple entities, make comparisons instead of expressing direct opinions, use a lot of sarcasm, etc.

Unsupervised versus Supervised Sentiment Classification
The sentiment classification task can be formulated as a supervised or an unsupervised classification problem, depending on whether there are known examples of documents belonging to positive or negative sentiments. Unsupervised sentiment classification involves the application of a sentiment lexicon of opinion-related positive or negative terms to evaluate text in the document. On the other hand, supervised sentiment classification applies machine-learning algorithms (such as support vector machines (SVMs) and neural networks) to textual feature representations to derive the relationships between features of the text segment and the opinions expressed in the document. In many practical situations, known class examples are created by experts who read the documents or use rules. Then, if a text review’s numeric rating is four or more stars, then the review is positive. If no known class examples are possible, then analysts have to use an unsupervised classification of sentiments.

Supervised classification is typically performed at the document level. Once enough labeled examples are available, any commonly used classification models can be trained, validated, and tested to check their performances. Published research on the topic of model performance in supervised classification is often based on product review data, which typically has a text review and an overall numeric rating on a scale of one to five stars. Often, a review rating of four to five stars is considered a positive rating, and a review rating of one to two stars is considered a negative rating. Many researchers have shown that the naïve Bayes classifier and (SVMs) perform reasonably well for the supervised classification task (Pang et al. 2008). The main challenge for modelers is to carefully select the inputs from text features such as terms and their frequencies (often weighted or normalized), part-of-speech tags, opinion lexicons (general, context-specific, and idioms), syntactic dependency (from parsing trees), and the handling of negation words (such as “not”).

Unsupervised classification is typically performed at the sentence level. There are two types of unsupervised classification methods: lexicon-based and syntactic-pattern based. The lexicon-based approach can be used for sentence- and aspect-level sentiment classification. It uses lexicons (opinion words) and involves identifying which opinion words are related to which attributes in a sentence. The relationships between opinion words and attributes are identified via dependency relationships obtained through parsing. For example, in the sentence, “The picture quality is outstanding,” the opinion word “outstanding” and the attribute “picture quality” share the same dependency relationship with the verb “is.” Consequently, the sentence is deemed to be expressing a positive sentiment about picture quality. If a clear dependency is not observed between an opinion word and an attribute, then how close an opinion word is to an attribute in a sentence can be used to judge the polarity of the attribute. This process can get very complex, depending on how long the sentence is, how many attributes are being mentioned in the same sentence, whether both positive and negative polarity words are used in the same sentence, whether negation is used, and so on. Once sentiment values are computed for each word-attribute combination, they are typically combined using appropriate normalization or weights to come up with an overall sentiment score. On the other hand, a syntactic-pattern-based approach involves defining part-of-speech tags and the keywords AND, OR, NOT, BUT, etc. Primarily useful in contextual analysis when performing phrase-level analysis, this method can be used to develop a variety of rules for better accuracy. For example, a simple pattern such as <subject> <NOT> <verb> can be used to extract negative phrases like, “This <feature> does <not> <verb> as advertised.” We delve deep into this approach in the rule-based models section later in this chapter.

SAS Sentiment Analysis Studio Overview

SAS Sentiment Analysis Studio is a comprehensive solution to the multifaceted challenges of analyzing sentiment in input documents. It enables users to automatically evaluate the opinions and feelings expressed about an entity (or object, person, event, or experience), either at the entity level or at the attribute (or feature) level. Generally speaking, SAS Sentiment Analysis Studio can use statistical models (such as naïve Bayes classifiers), rule-based models (using advanced linguistic technologies), or a combination of statistical and rule-based (hybrid) models.

Here is the architecture of SAS Sentiment Analysis Studio:
The training corpus typically consists of documents that have known class examples of sentiments (such as positive, negative, neutral, and unclassified). These are often compiled by domain experts. The software first extracts linguistic features from the training corpus. Then, using those linguistic features as inputs and known class examples as targets, the software trains and validates several statistical models and identifies the best model. The linguistic features are used by subject matter experts to write linguistic rules that are used in training and validating rule-based models. Hybrid models are created and tested by combining rule-based and statistical models. Any of the developed models can be used to score new documents.

If you also have a license for SAS Sentiment Analysis Workbench, you can use the software to create visually appealing graphs and charts to track sentiment over time and to customize search or drill down to explore sentiments at a very granular level.