Chapter 8
Conducting Sentiment Analysis and Post-Sentiment Data Exploration through Automated Means

Shalin Hai-Jew
Kansas State University, USA

ABSTRACT

One new feature in NVivo 11 Plus, a qualitative and mixed methods research suite, is its sentiment analysis tool; this enables the autocoding of unlabeled and unstructured text corpora against a built-in sentiment dictionary. The software labels selected texts into four categories: (1) very negative, (2) moderately negative, (3) moderately positive, and (4) very positive. After the initial coding for sentiment, there are many ways to augment that initial coding, including theme and subtheme extraction, word frequency counts, text searches, sociogram mapping, geolocational mapping, data visualizations, and others. This chapter provides a light overview of how the sentiment analysis feature in NVivo 11 Plus works, proposes some insights about the proper unit of analysis for sentiment analyses (sentence, paragraph, or cell) based on text dataset features, and identifies ways to further explore the textual data post-sentiment analysis—to create coherence and insight.

INTRODUCTION

In a complex multimedia environment, the most common base form of data is still textual. Even common multimedia—audio and video—tend to have a textual component (the transcript) that captures some of the informational value of the digital file; likewise, imagery often have "alt-text" and metadata descriptors. In text may be found residuals of common human endeavors: inter-communications, record-keeping, research, and history-keeping. In this current age of social media, much of the outpouring of created data is in text form. In a time of informational plenty (in terms of certain topics), there are computational ways of "read" and "understand" information from various text sets.

DOI: 10.4018/978-1-5225-0648-5.ch008

The inherent structures of language and their common use for the encoding and decoding of shared interchanges and understandings have made it possibly to apply computational means (in scalable ways and at machine speeds) to extract insights from texts and text corpora (collections of texts). One capability involves classifying texts into sentiment polarities or valence (either a binary positive or a binary negative) or into sentiment categories (various gradations of sentiment based on both polarity and intensity/arousal). Sentiment analysis (also known as "opinion mining") is the identification of attitude, whether positive or negative, towards a particular issue, event or phenomenon, concept, product, service, person, organization, or another element. How sentiment is understood computationally varies, but one of the simplest approaches involves using a pre-defined sentiment dictionary or word set and comparing the terms of a target text or text corpus against that and classifying the target text based on identified sentiment. The number of opinion words found informs the sense of intensity of sentiment in that text corpus. Technically, the practice of sentiment analysis stems from work in natural language processing (NLP), artificial intelligence (AI), computational linguistics, data mining, and text analysis.

Sentiment or opinion analysis may be applied to a text and then the coded texts may be further machine queried and processed for insight. Themes and subthemes may be extracted to capture a summary view of what a text or a text corpus is "about." Word frequency counts are used as a proxy measure of the focus or emphasis of a particular text corpus. The more frequently words that are listed, the more mentions are being represented, and the greater interest there is in the group of documents around that topic. This approach has been applied to Tweetstream texts, #hashtag conversations texts, and formal article sets in particular domain fields, among other applications. Text searches may be applied to see every instantiation of selected words, n-grams (consecutive verbatim text sequences), phrases, names, formulas, or other textual elements…and the proximity text leading up to- and away- from the selected text presented in a word tree to show the contextual gist of the phrases. Content and messaging data that people share via social media platforms are thought to be highly opinion-laden data: think upvotes and downvotes, likes and dislikes, votes about hot-or-not, reviews about everything bought or sold (and the sellers besides), raging feuds, and people calling out each other. Data extracted from "social media"—broadly including email systems, social networking sites, content-sharing sites, microblogging sites, blogging sites, crowd-sourced encyclopedias, and others—may be depicted in sociograms (social networks displayed as graphs). Also, it is possible to capture textual data with various locational markers and to map those locations onto a two-dimensional geolocational map.

There are two basic hypotheses in this work:

Hypothesis 1: There is an optimal unit of analysis for sentiment analysis (sentence, paragraph, or cell) using the NVivo 11 Plus tool based on the dataset type (and how the textual data is structured in that corpus).

Hypothesis 2: There are data analytics benefits to defining a sequence of data analytics steps post-sentiment analysis to add value to the coded data (using NVivo 11 Plus).

One of the new features of NVivo 11 Plus, a qualitative and mixed methods research suite, is its sentiment analysis tool which enables the autocoding of text corpora against a sentiment dictionary (a listing of words and lexical elements which are seen to contain inherent sentiment or attitude). The software labels texts into four categories: very negative, moderately negative, moderately positive, and very positive. After the initial coding for sentiment, there are still a number of ways to augment that initial coding by using various query and related data visualization features in NVivo 11 Plus, including