Big Data Sentiment Analysis of Customers' Reviews, and the Prediction of Companies' Stock Performance

TREO Talk Paper

Sumali Conlon
University of Mississippi
sconlon@bus.olemiss.edu

Chad Diaz II
University of Mississippi
cdiaz@bus.olemiss.edu

Abstract

A great deal of research has been done on the prediction of stock prices, using various techniques to analyze data from several sources. Among the many papers in this area, Schumaker and Chen (2009) use a predictive machine learning approach to analyze financial news articles to predict stock prices. Kim et al. (2014) analyze Korean Language news articles to predict rises and falls of the KOSPI (Korean Composite Stock Price Index). Finally, Bollen and Mao (2011) and Mittal and Goel (2012) analyze collective mood states derived from large-scale Twitter feeds to see whether these are correlated with the value of the Dow Jones Industrial Average (DJIA) over time.

Among the many factors that may cause stock prices to change over time, we are interested in finding out whether or not customers’ opinions about the products they purchased from those companies can be used to predict the companies’ stock performance. For example, if certain products are very successful and customers like them a lot, do their comments help to predict whether the stock prices rise for the companies that produce those products? We use sentiment analysis techniques, consisting of several machine learning algorithms, such as linear regressions, Naïve Bayes, support vector machines (SVM), and neural networks, to analyze a targeted portion of Amazon’s product reviews. This corpus consists of 34,686,770 reviews between June 1995 and March 2013, available at the Stanford Large Network Dataset Collection (SNAP) web site. A company’s stock performance is predicted by measuring customer opinions about its products, whether they are satisfied with those products or not. The satisfaction values returned by the algorithms are tested to see whether they can be used to predict the companies’ stock performance by computing p-values using time series techniques, in particular, Granger Causality tests.

References


