

News, Sentiment, and Financial Markets: A Computational System to Evaluate the Influence of Text Sentiment on Financial Assets.

Stephen Kelly

Abstract

With the advent of the internet and digitisation of news and books, the volume of unstructured text has increased dramatically in recent years. This deluge of information is set to grow and come from new and unconventional sources. New and innovative techniques and tools for manipulating this data and making sense of it will become essential. The work in this thesis consists of a system that analyses the content of news, extracting a sentiment time series variable, and uses this variable in a time series modelling component to determine any inter-relationships between changes in the price of financial assets. Each component of the system attempts to remove human subjectivity from the modelling process to allow the system to compute a sentiment variable and investigate its statistical significance and explanatory power by employing several time series models. The system includes a number of processes and components to achieve this goal such as data harvesting, processing, text analysis, time series modelling, hypothesis testing, and visualisation. The work described in this thesis includes contributions to the area of text and content analysis, information retrieval, time series analysis, statistical and econometric modelling.

A number of methods studied in the literature have incorporated text data with financial analysis and prediction models. A review of some of the main studies and systems that have combined methods from each discipline is presented in Chapter 2. Chapter 3 describes the system developed in this thesis and its capabilities. The system is evaluated for different data inputs in Chapter 4, where the influence of different news types and sources is investigated for equity and commodity markets. The final chapter concludes the thesis by summarising the contributions and outlining future work.

This thesis represents a combination of work that contributes to the areas of content analysis and financial and statistical modelling. The main contribution of this thesis is in the implementation and evaluation of a system that incorporates methods from text analysis and time series modelling. The novelty of the system lies in the ability to compute a time series from text that acts as a proxy for sentiment that can be aggregated with financial time series data in a statistical model to estimate the impact of news on the financial asset. An evaluation of this system is presented where the explanatory power of the sentiment variable for financial returns is investigated. It is shown that the news source and text type play an important role when computing a proxy for sentiment that has statistically significant explanatory power for financial returns. The time varying influence of sentiment on financial returns is noted with a relationship made between economic business cycles and volatility. The system is evaluated using data from two financial markets, the equity and commodities markets, and in both instances it is found that a proxy for sentiment extracted from news has a statistically significant influence on financial returns.