Abstract

The growth of the media in terms of its expansion across new technological mediums has greatly increased the availability of information. Specifically in the business world, this has translated into greater access to news and detailed reports relating to the world’s largest corporations. Such information generally conveys some sort of indication of predicted success or failure of the companies. This means that, more than ever, those trading stocks are reading and likely being influenced by several news sources every day. We believe that this influence can push traders to make stock trades that they might not otherwise have made. Thus, this work is intended to show that there exists a correlation between the information released by major news sources and the change in stock price immediately following the release of such information. Additionally, we seek to rank news organizations based on the amount of influence our model predicts that they have on stock prices.

1 Introduction

This goal of this project is to find a correlation between news sites and the effect of their securities analyses/opinions on the value of the securities being analyzed. The project varies from many existing projects that seek to predict the outcome of the stock market based on the vocabulary related to a particular security or current work in neuroeconomics[6] [4] or behavioral finance[8]. Instead, the work addressed in the proposal seeks to determine whether the major news outlets have an effect on stock prices, and if so, how reliable or strong the association is. The metric associated with this correlation is discussed in the following sections.

The technologies utilized in this project are many. We will be using the Java programming language to develop each part of our project. We will be using multiple techniques to acquire data (stock prices and news articles), and also to analyze the data (article rating - i.e., how strongly it speaks for or against the company being analyzed). We will be doing text analysis to obtain the article
rating, and exploring different types of text analysis - phrase-based, word map, etc. - to procure the best results.

The below sections offer more detail on the proposed implementation of our project. The Proposed Solution details the different "atomic" parts and algorithms required to accomplish our task; the Data/Datasets show the specific members of the two different pieces of data - articles and stock prices - upon which our analysis relies.

2 Problem Statement

Our goal is to determine if there is a correlation between news article reports on stock prices and the pricing of stocks on the stock market.[1] [9] If there is a correlation, we will determine which news source is the best predictor of the stock market prices.

3 Proposed Solution

In order to determine if a correlation is present, we will build a software application that will examine the past year’s worth of online news articles about a given company and compare that information to actual stock market data. After all of our comparisons are complete, we will can then determine which news source is the most accurate stock price predictor. To accomplish these tasks, we have separated our solution into multiple components. Each component will work as shown below:

3.1 Article Extraction

The first step is to collect news articles from various news sources via the Internet. To minimize the different sites required to gather data, Google News is used as an aggregator of articles from different news sources. Given the name of a company to search, Google News is queried and a list of articles is then extracted and inserted into our MySql database. This greatly reduces the time required in the follow steps since all articles are cached locally. Once all the articles have been extracted, they are in the correct form to be rated.

3.2 Article Rating

The next step that articles take after being extracted from a news source is pass through an Article Rater. This process assigns each article an Article Company rating (or ACR). This score, represented on a scale of -1 to 1, indicates how positively or negatively the article portrays the company that is the subject of the article. Since we believe that the results of the Article Rater will be vitally important to getting good overall results, we have designed our application to make this component pluggable. In this way, we can write several different algorithms that compute an ACR and easily choose between them at runtime.
We will be able to use this pluggability to compare the performance of our system across different algorithms.

We are currently looking at several methods for assigning an ACR to each article. These methods all fall under the subsection of data mining called sentiment mining or sometimes called opinion mining. These techniques involve part of speech tagging, correlation between search query results, bayesian filtering, and other classification techniques. Most of the work in these areas have been using the concrete example of product reviews. Some of our work will involve finding ways that the algorithms need to be tuned and tweaked to get good results on news data. [7, 5]

3.3 Stock Price Extraction

Procuring historical stock prices off of the Internet and storing them in our own database for quick access. This process simply accesses a Google Finance URL that returns historical stock prices of a certain company and persists them to a local database for quick, Internet-less access in the future. The process can be repeated for any number of companies and any number of dates. Google returns a comma-delimited text file (.csv) for each company and set of dates that is easily parsed and committed to a database via a few Java ODBC methods. The specific dataset obtained in Stock Price Extraction is detailed in section 4.2

3.4 Article Prediction Scoring

The relevance of the article rating to the performance of the stock the following day. We use the same -1 to 1 scale that is used in the Article Rating for consistency. This section effectively gives a grade to the "prediction" weaned from article analysis. That is, if an article rated a stock 1.0 ("perfect"), and the stock price the following day went up 10 percent (we are using 10 percent or more as a "perfect" stock value increase), then we would give the Article Rating a score of 1.0 because it was perfectly correlated with the performance of the stock. Likewise, if the article gave a security a 1.0, and the stock went down 10 percent the next day, then the rating would receive a score of -1 because the stock performance was exactly opposite of what it predicted.

3.5 Source Influence Score Summary

Once the source influence scoring has occurred, the score must be grouped in order to determine the news source’s influence on a company’s stock price. One way to achieve this goal is to simply average all of the source influencing scores for a particular news source.

Another benefit of this method of calculating the influence score for all company-news source pairs is that other statistics can also be generated. They are described in the following list.

1. The influence each news source has in general over all companies (the main calculation)
2. How companies are influenced over all news sources
3. How much a particular company is influence from a particular news source

4 Data/Datasets

Since our project deals with both news articles and the stock market, we are creating extracting datasets from information freely on the web. The following sections describe these datasets in more detail.

4.1 News Article Dataset[3]

Google news provides an aggregation of multiple news sources in one location. Unfortunately, Google does not offer any API's to programmatically access their information. Web scraping techniques are instead used to extract the news information into one organized database. The dataset is generated based on the following parameters:

- List of companies for which the articles should relate
- Beginning/end date for article publication
- Max number of articles to find

The dataset contains the following fields for a given article:

- article_id : int(11)
- news_source : varchar(1000)
- headline : varchar(1000)
- text : text
- url : varchar(1000)
- publication_date : date
- company : varchar(1000)

4.2 Stock Values[2]

Google Finance provides an easily parsed .csv file to access historical security (stock) data. However, this is only accessible one company at the time, and only with a connection to the Internet. We want to be able to access our data without a connection, and also to access only the data desired for the task at hand. For this reason, the stock values are persisted to a local database. The dataset is generated based on the following parameters:
• Stock prices for companies we want to analyze
• Beginning/end dates for stock quotes

The dataset contains the following fields for a given company stock:

• company : varchar(20)
• symbol : varchar(20)
• market : varchar(20)
• date : date
• opening : float
• high : float
• low : float
• close : float

Where company is the name of the company whose stock we are analyzing, symbol is the stock ticker symbol for the company, market is the stock market where the security can be found (NYSE, NASDAQ, etc.), opening is the opening price of the stock for the date, high is the high price of the stock for that day, low is the low price of the stock for that day, and close is the price at which the stock closed on that date.

References


Figure 1: Our KICK ASSSS DESIGN