Chapter 53
Using Online Data in Predicting Stock Price Movements:
Methodological and Practical Aspects

František Dařena
Mendel University in Brno, Czech Republic

Jonáš Petrovský
Mendel University in Brno, Czech Republic

Jan Přichystal
Mendel University in Brno, Czech Republic

Jan Žižka
Mendel University in Brno, Czech Republic

ABSTRACT

A lot of research has been focusing on incorporating online data into models of various phenomena. The chapter focuses on one specific problem coming from the domain of capital markets where the information contained in online environments is quite topical. The presented experiments were designed to reveal the association between online texts (from Yahoo! Finance, Facebook, and Twitter) and changes in stock prices of the corresponding companies. As the method for quantifying the association, machine learning-based classification was chosen. The experiments showed that the data preparation procedure had a substantial impact on the results. Thus, different stock price smoothing, the lags between the release of documents and related stock price changes, levels of a minimal stock price change, different weighting schemes for structured document representation, and classifiers were studied. The chapter also shows how to use currently available open source technologies to implement a system for accomplishing the task.

DOI: 10.4018/978-1-7998-9020-1.ch053

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

A lot of research has been focusing on incorporating the vast amount of data available online into models of various social and economic phenomena in miscellaneous fields of science. The data, which is generated not only by domain experts but also by regular people, can provide new perspectives and potentially complementary information to conventional quantitative and objective evidence (Kearney & Liu, 2014). The data often comes from non-traditional and contemporary information sources and environments, like social networks or microblogging sites. The data typically has a form of unstructured texts that are published by different types of subjects, without the time and spatial limits.

On one hand, there are many logical reasons for this trend: there is a lot of data freely available, many different sources can be combined together, the data is often accessible immediately, in real time, and the variety of data items, opinions, and perspectives can be great. On the other hand, collecting, storing, and processing such data relate to several problems: tools for collecting the data need to be available, it is necessary to handle multiple systems and data formats, it is necessary to consider differences in data quality, reliability, objectivity etc.

One of the domains, where using online data is very topical and plays an important role in analyzing the studied systems, is the field of capital markets. Here, the data provided by digital media can help, e.g., in explaining less rational factors such as investors' sentiment or public mood as influential for asset pricing and capital market volatility (Bukovina, 2016).

One of the greatest advantages of using online resources for decision making support in the domain of capital markets is the timeliness of the information, which is particularly important for investment decisions. The quality of the messages posted in online environments (such as microblogs or discussions in social networks) is, however, generally low. That is why Internet postings have been the least frequently studied source of textual sentiment (Kearney & Liu, 2014). Despite all difficulties, content generated by web users has become a widely accepted resource for determining sentiment or opinions related to different aspects of the public mood (Tumasjan et al., 2011). It has been also shown that a large number of people participating in a content generation process enables the creation of artifacts that are of equal or superior quality than those made by experts in the respective field (Gottschlich & Hinz, 2014). Messages from millions of people are also unlikely to be biased (Mostafa, 2013).

There exist numerous studies focusing on the usefulness of textual data for predictions related to stock prices. Majority of them focused on an aggregate level (e.g., the level of a stock market index), worked with a single source of texts (e.g., newspaper articles or financial reports), or required additional expert knowledge (e.g., a list of words or expressions that are usually related to positive and negative stock price movements).

In our work, we focus on analysis at the micro level, namely at the level of individual companies. The goal is to determine whether there are some associations between the content of online texts related to a company and the movements of the stock prices of that company. We also wanted to avoid the necessity of using various lexicons provided by domain experts, which might perform poorly in previously unknown situations (Eisenstein, 2017). In our research, we also combine documents from three different sources, Yahoo! Finance, Facebook, and Twitter collected over a period of about 8 months. A machine learning-based approach is applied in order to find out whether the content plays an important role in revealing the document-stock price movement association.

Section Background gives a brief overview of the current state of research in the field of capital markets where text data is used as one of the possible sources of data. The Experimental Procedure Background

26 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/using-online-data-in-predicting-stock-price-movements/283017

Related Content

ICTs: Convenient, Yet Subsidiary Tools in Changing Democracy
Kerill Dunne (2015). *International Journal of E-Politics (pp. 1-13).*
www.irma-international.org/article/icts/127686

The Use of Twitter During the 2013 Protests in Brazil: Mainstream Media at Stake
Nina Fernandes dos Santos (2020). *Handbook of Research on Politics in the Computer Age (pp. 181-202).*
www.irma-international.org/chapter/the-use-of-twitter-during-the-2013-protests-in-brazil/238224

A Modern Socio-Technical View on ERP-Systems
Jos Benders, Ronald Batenburg, Paul Hoekenand Roel Schouteten (2009). *Handbook of Research on Socio-Technical Design and Social Networking Systems (pp. 429-439).*
www.irma-international.org/chapter/modern-socio-technical-view-erp/21424

Nominalizations in Requirements Engineering Natural Language Models
Claudia S. Litvak, Graciela Dora Susana Hadadand Jorge Horacio Doorn (2019). *Advanced Methodologies and Technologies in Media and Communications (pp. 192-202).*
www.irma-international.org/chapter/nominalizations-in-requirements-engineering-natural-language-models/214552

“I’ve Got a Situation and Would Appreciate Your Experience”: An Extra-Organizational Virtual Community of Practice for Independent Professionals
Enrique Murillo (2012). *International Journal of Virtual Communities and Social Networking (pp. 52-80).*
www.irma-international.org/article/got-situation-would-appreciate-your/75779