

Construction and Application of Sentiment Lexicons in Finance

Kazuhiro Seki, Konan University, Kobe, Japan

Masahiko Shibamoto, Kobe University, Kobe, Japan

ABSTRACT

This article proposes an approach to constructing sentiment lexicons in the financial domain. The approach takes advantages of news bulletins and a given financial variable, such as stock prices, to generate candidates of sentiment expressions by fusing the two data sources. The candidates are then filtered based on their co-occurrences with financial seed words and are subsequently expanded by analogical reasoning using distributed representation of words. Evaluative experiments on real-world news and stock price data shows that the resulting lexicons are mostly reasonable and capture the characteristics of the target financial variables. As a potential application, trading simulation is also carried out based on the resulting financial sentiment lexicons, demonstrating the utility of the lexicons.

KEYWORDS

Distributed Representation, Sentiment Analysis, Stock Market, Text Mining

1. INTRODUCTION

In the last decades, Consumer Generated Media (CGM), such as microblogs, customer reviews, and Q&A forums on the World Wide Web (WWW), gained much popularity and have been increasingly used all over the world. The information posted on those media are valuable and often have strong influence on our daily decisions (e.g., choosing products to buy or places to visit). To make the most of the rich information, there has been much research for mining textual data on the vast WWW (Imran et al., 2015).

One of the main themes in text/web mining is sentiment analysis (Liu, 2015), which generally estimates the sentiment of an input text. In analyzing sentiment, sentiment lexicons are often utilized as one of the essential linguistic resources. A simplest form of sentiment lexicons is compiled as two lists of words or expressions; one containing “positive” ones, such as “good” and “excellent”, and the other containing “negative” ones, such as “bad” and “terrible”. In addition, sentiment lexicons may have a polarity score associated with each expression, which indicates how strong the sentiment of the expression is. As an example, Table 1 presents a fragment of sentiment lexicons, SentiWordNet (Baccianella et al., 2010), showing some positive and negative words with their polarity scores. Note that sentiments types are not necessarily limited to the dichotomous positive/negative and more fine-grained sentiment lexicons may have others types of sentiments (e.g., *ashamed*, *scared*, *excited* and *relieved*) (Takamura et al., 2005).

Text mining has been also applied to the financial domain. For example, many researchers analyzed news articles to predict a financial variable (e.g., stock prices), where their approaches can be categorized into two types. One is to use words appearing in news articles as explanatory variables and the target financial variable as the explained variable and directly models their relations by

DOI: 10.4018/IJMDem.2018010102

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Table 1. A fragment of SentiWordNet (Baccianella et al., 2010)

Positive		Negative	
Term	Score	Term	Score
Splendid	1.000	Scrimy	1.000
Superb	0.875	Unfortunate	0.889
Valued	0.750	Upset	0.875
Quintessential	0.625	Villainous	0.778

regression models (Schumaker and Chen, 2009). The other takes a deeper look at news articles and finds words or phrases which would positively/negatively affect the target financial variable, where precompiled lists of positive and negative words/phrases are needed (Li et al., 2014). The latter type of approaches is similar to sentiment analysis using sentiment lexicons in essence.

As a first step toward financial text mining, this study aims at constructing financial sentiment lexicons composed of positive/negative expressions that have a short-term impact on a given financial variable. Here, it should be emphasized that the resulting sentiment lexicons are dependent on the target financial variable since sentiment expressions are expected to be different from variables to variables. For example, stock prices of hospitals and drug makers would not be affected in the same way by the news reporting the failure of repealing the Affordable Care Act (a.k.a. Obamacare).¹ To this end, our approach looks at the relation between the movement of financial variable and news headlines in order to construct an initial sentiment lexicons, which are then expanded by analogical reasoning by distributed representation (Mikolov et al., 2013a). Further, as a possible application of the resulting sentiment lexicons, they are used for judging the sentiment of news headlines, which is then used for trading simulation.

The contributions of the present work can be summarized as follows:

- A systematic framework is proposed to identify candidates of sentiment terms based on news articles and a given financial variable, which results in variable-dependent sentiment lexicons.
- Initial lexicons are automatically constructed and are further expanded by analogical reasoning using distributed representation, where synonyms and antonyms are distinguished.
- The validity of the proposed approach is shown by experiments on real-world data.
- The utility of the resulting lexicons is demonstrated by trading simulation.

The remaining of the paper is structured as follows: Section 2 summarizes the related work on constructing sentiment lexicons in the general domain and the financial domain. Then, Section 3 describes our approach to constructing financial sentiment lexicons, and Section 4 reports on the results of evaluative experiments. Finally, Section 5 provides a brief summary and possible future directions.

2. RELATED WORK

There has been much research on automatic construction of sentiment lexicons. The following subsections summarize the related work and then introduce some attempts carried out specifically targeting the financial domain.

2.1. Automatic Construction of Sentiment Lexicons

The previous work on automatic construction of sentiment lexicons can be categorized into two approaches. One is dictionary-based and the other is corpus-based. The dictionary-based approach

12 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/article/construction-and-application-of-sentiment-lexicons-in-finance/196247

Related Content

Intelligent Personalization Agent for Product Brokering

Sheng-Uei Guan (2009). *Encyclopedia of Multimedia Technology and Networking, Second Edition* (pp. 703-709).

www.irma-international.org/chapter/intelligent-personalization-agent-product-brokering/17469

Multimedia Social Network Modeling using Hypergraphs

Giancarlo Sperli, Flora Amato, Vincenzo Moscato and Antonio Picariello (2016). *International Journal of Multimedia Data Engineering and Management* (pp. 53-77).

www.irma-international.org/article/multimedia-social-network-modeling-using-hypergraphs/158111

Mobile Multimedia Collaborative Services

Do Van Thanh, Ivar Jørstad and Schahram Dustdar (2009). *Handbook of Research on Mobile Multimedia, Second Edition* (pp. 931-945).

www.irma-international.org/chapter/mobile-multimedia-collaborative-services/21054

The Haunted School on Horror Hill: A Case Study of Interactive Fiction in an Elementary Classroom

Kevin Kee, Tamara Vaughan and Shawn Graham (2011). *Gaming and Simulations: Concepts, Methodologies, Tools and Applications* (pp. 1405-1416).

www.irma-international.org/chapter/haunted-school-horror-hill/49457

Movie Video Summarization- Generating Personalized Summaries Using Spatiotemporal Salient Region Detection

Rajkumar Kannan, Sridhar Swaminathan, Gheorghita Ghinea, Frederic Andres and Kalaarasi Sonai Muthu Anbananthen (2019). *International Journal of Multimedia Data Engineering and Management* (pp. 1-26).

www.irma-international.org/article/movie-video-summarization--generating-personalized-summaries-using-spatiotemporal-salient-region-detection/245751