Expert Systems with Applications 42 (2015) 306–324

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment

Arman Khadjeh Nassirtoussi [a,*], Saeed Aghabozorgi [a], Teh Ying Wah [a], David Chek Ling Ngo [b]

[a] Department of Information Science, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[b] Research & Higher Degrees, Sunway University, No. 5, Jalan University, Bandar Sunway, 46150 Petaling Jaya, Selangor DE, Malaysia

ARTICLE INFO

Article history:
Available online 10 August 2014

Keywords:
News mining
News semantic analysis
Market sentiment analysis
Market prediction
FOREX prediction

ABSTRACT

In this paper a novel approach is proposed to predict intraday directional-movements of a currency-pair in the foreign exchange market based on the text of breaking financial news-headlines. The motivation behind this work is twofold: First, although market-prediction through text-mining is shown to be a promising area of work in the literature, the text-mining approaches utilized in it at this stage are not much beyond basic ones as it is still an emerging field. This work is an effort to put more emphasis on the text-mining methods and tackle some specific aspects thereof that are weak in previous works, namely: the problem of high dimensionality as well as the problem of ignoring sentiment and semantics in dealing with textual language. This research assumes that addressing these aspects of text-mining have an impact on the quality of the achieved results. The proposed system proves this assumption to be right. The second part of the motivation is to research a specific market, namely, the foreign exchange market, which seems not to have been researched in the previous works based on predictive text-mining. Therefore, results of this work also successfully demonstrate a predictive relationship between this specific market-type and the textual data of news. Besides the above two main components of the motivation, there are other specific aspects that make the setup of the proposed system and the conducted experiment unique, for example, the use of news article-headlines only and not news article-bodies, which enables usage of short pieces of text rather than long ones; or the use of general financial breaking news without any further filtration.

In order to accomplish the above, this work produces a multi-layer algorithm that tackles each of the mentioned aspects of the text-mining problem at a designated layer. The first layer is termed the Semantic Abstraction Layer and addresses the problem of co-reference in text mining that is contributing to sparsity. Co-reference occurs when two or more words in a text corpus refer to the same concept. This work produces a custom approach by the name of Heuristic-Hypernyms Feature-Selection which creates a way to recognize words with the same parent-word to be regarded as one entity. As a result, prediction accuracy increases significantly at this layer which is attributed to appropriate noise-reduction from the feature-space.

The second layer is termed Sentiment Integration Layer, which integrates sentiment analysis capability into the algorithm by proposing a sentiment weight by the name of SumScore that reflects investors' sentiment. Additionally, this layer reduces the dimensions by eliminating those that are of zero value in terms of sentiment and thereby improves prediction accuracy.

The third layer encompasses a dynamic model creation algorithm, termed Synchronous Targeted Feature Reduction (STFR). It is suitable for the challenge at hand whereby the mining of a stream of text is concerned. It updates the models with the most recent information available and, more importantly, it ensures that the dimensions are reduced to the absolute minimum.

The algorithm and each of its layers are extensively evaluated using real market data and news content across multiple years and have proven to be solid and superior to any other comparable solution. The proposed techniques implemented in the system, result in significantly high directional-accuracies of up to 83.33%.

On top of a well-rounded multifaceted algorithm, this work contributes a much needed research framework for this context with a test-bed of data that must make future research endeavors more convenient. The produced algorithm is scalable and its modular design allows improvement in each of its layers in future research. This paper provides ample details to reproduce the entire system and the conducted experiments.

* Corresponding author.
E-mail address: armankhnt@gmail.com (A. Khadjeh Nassirtoussi).

http://dx.doi.org/10.1016/j.eswa.2014.08.004
0957-4174/

1. Introduction

Today's biggest economies of the world are market-economies. With markets being the heart of economies, it is paramount to understand, utilize and predict them to the betterment of society. At the core of every market lie supply–demand equilibriums. Market participants provide supply or demand into the markets based on their perception of the world. Human perception is limited to the information available. Information is made available constantly via news channels. Hence, undeniably news content has an impact on market-movements. However, a more granular quantification of this relationship between markets and the news has been extremely challenging, because news contains unstructured information in form of language. One approach to address unstructured data and extract structured data from it is the development of specialized search engines like the financial news semantic search engine by Lupiani-Ruiz et al., 2011. However, a search engine like the above is limited to extracting the available numeric data in the texts. Deciphering language by machine constitutes the complex field of natural language processing (NLP). From this perspective this work lies at the intersection of NLP and opinion mining or sentiment analysis which are recently being increasingly researched for many emerging needs (Cambria, Schuller, Yunqing, & Havasi, 2013). There have been some early-stage efforts to make stock-market related predictions based on news-text (Hagenau, Liebmann, & Neumann, 2013; Mittermayer, 2004; Schumaker, Zhang, Huang, & Chen, 2012; Tetlock, Saar-Tsechansky, & Macskassy, 2008; Wuthrich et al., 1998) and very few Foreign-Exchange-Market (FOREX) related ones (Peramunetilleke & Wong, 2002). However, there are some similarities between the two problems. When dealing with news and market-movements the basic strategy can be to try to draw a statistical relationship between the appearance of words and the market movements. In this scenario most words are representing themselves as features in a feature-vector matrix and the technique is termed as Bag-of-Words (Mahajan, Dey, & Haque, 2008; Schumaker et al., 2012; Wuthrich et al., 1998). Bag-of-Words has been widely used in many of the related works to markets and news and its primary downside is the huge number of features that it produces, which very easily leads to the curse-of-dimensionality (Pestov, 2013). Moreover, many words may represent the same idea, concept or thing and it may make a lot more sense to somehow have them abstracted accordingly. Thereby, proposing solutions to the above two challenges, namely, the latter or feature-selection in an abstracted form and the former, a way to tackle the curse-of-dimensionality via a feature-reduction technique are center-pieces of this work.

There are two main areas of contributions made in this work. A brief summary is provided below:

A – Proposal of novel text-mining methods in 3 areas:

1. Semantic-Abstraction and Integration via a novel feature-selection technique, termed, Heuristic-Hypernyms.
2. Sentiment Integration via a novel sentiment-weighting mechanism, termed, SumScore.
3. Dimensionality Reduction via a novel feature-reduction technique, termed, Synchronous Targeted Feature-Reduction.

B – Exploration of a specific novel use-case, namely: "Short-term FOREX prediction based on news-headlines".

1. Exploration of a new market-type through predictive text-mining. The predictive text-mining of news has not been explored before in the FOREX market to the best of our knowledge.
2. Usage of news article-headlines rather than news article-bodies. News-headlines have been explored in extremely few market-predictive text-mining research works before.
3. A novel solution to enable short-term prediction of 1 to 3 h after news-release.

All of the above 6 items are contributions of this work. The hybrid of the above new techniques produces results that are significantly higher compared to scenarios without them. A directional accuracy as high as 83.33% is achieved in experiments conducted on Euro/USD currency-pair intraday-movements which is laid out and discussed in detail later in this paper.

In the rest of this paper these sections follow: 2 – Literature review; 3 – Problem description; 4 – System description; 5 – Experimental results and evaluation; 6 – Concluding remarks and future research.

2. Problem description

The specific problem that this research addresses and its requirements is briefed in the below.

The first aspect of the problem definition is a focus on a specific market-type. In general, there are multiple types of financial markets, namely: 1 – Capital markets (Stock and Bond), 2 – Commodity markets, 3 – Money markets, 4 – Derivative markets, 5 – Future markets, 6 – Insurance markets and 7 – Foreign exchange markets. As their names imply, different assets are traded in each market; therefore they demonstrate different behaviors and separate research is conducted on each of them. As it is pointed out more specifically in the literature review section of this work, most of the works in the literature concerning some kind of usage of text-mining for a predictive purpose in a financial market is mostly attending to the stock-markets and specific company stocks based on textual content about those companies. Hence, this work enters a less explored financial market namely the foreign exchange market (FOREX) which facilitates the trading of currencies.

Furthermore, this work aims to take into use uncategorized breaking news rather than categorized news based on topic or company, etc. As pointed out in the literature the usual explored path in the past works is to isolate company-specific news, for example, and make predictions for the stock of a company based on that. However, the news channel that is used for this experiment is for financial breaking news. A focus on financial breaking news rather than a source of news that has all kinds of news pieces released is assumed to provide logical relevance and avoid noise. This is inspired by what traders in financial markets actually read. But no further categorization of news is utilized.

Moreover, in terms of the length of text, subject of this research is short-texts of news-headlines. The requirement of using news article-headlines rather than news article-bodies creates a text-mining focus on short texts rather than long texts for the proposed system. Naturally, when short pieces of texts are concerned there is less repetition of words in the same document and there are also fewer irrelevant words. Therefore, in such a context the level of significance of a word in a news piece cannot be determined by its repetition within it; however, at the same time there is less noise in the space as headlines are usually concise.

In terms of prediction time-line in the financial markets, both short and long-term predictions are subjects of research. In this work, however, the short-term prediction is explored as sudden impacts of news on the market are of interest and with the passage of time the number of factors producing noise on the initial impact increases. The short-term prediction that targets market-moves within the same day is termed as intra-day market prediction. To be specific, what is predicted is the directional movement (Up or Down) of the market (price of a currency pair e.g. EUR/USD) 1-h after the end of a 2-h interval which includes the news-headlines released within it. This upwards or downwards movement at the