Forex-Foreteller: Currency Trend Modeling using News Articles

Fang Jin, Nathan Self, Parang Saraf, Patrick Butler, Wei Wang, Naren Ramakrishnan
Department of Computer Science,
Discovery Analytics Center,
Virginia Tech, Blacksburg, VA 24061
{jfang8, nwself, parang, pabutler, tskatom, naren}@cs.vt.edu

ABSTRACT

Financial markets are quite sensitive to unanticipated news and events. Identifying the effect of news on the market is a challenging task. In this demo, we present Forex-foreteller (FF) which mines news articles and makes forecasts about the movement of foreign currency markets. The system uses a combination of language models, topic clustering, and sentiment analysis to identify relevant news articles. These articles along with the historical stock index and currency exchange values are used in a linear regression model to make forecasts. The system has an interactive visualizer designed specifically for touch-sensitive devices which depicts forecasts along with the chronological news events and financial data used for making the forecasts.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition; Parameter learning*

General Terms

Economics, Design

Keywords

Topic discovery, currency markets, sentiment analysis.

1. INTRODUCTION

Foreign market exchanges are among of the most liquid financial markets in the world. According to the Bank of International Settlements, the average daily turnover is estimated at $3.98 trillion as of April 2010 with a growth of 20% as compared to April 2007. Traders include governments, financial institutions and retail investors. Currency exchange rates are the most important aspect of international trade and investments. In an increasingly challenging and competitive market, having an estimate of currency movement is a holy grail. However, the irregular effects of economic, political and environmental factors make currency market forecasting a very complex task. Many traders choose to hedge their currency investments due to such difficulties.

In this demo, we present a novel system for modeling currency exchange rates, which we call Forex-foreteller [1]. Forex-foreteller not only makes potential forecasts and generates warnings but also allows traders to view the chronological set of news events and historical financial data used in making forecasts. A visualization has been specifically designed for a large touch screen device. Currently, the system forecasts currency exchange rates for four Latin American countries: Argentina, Brazil, Chile, and Columbia, and could easily be extended for other markets, given comparable (or greater) data coverage.

2. RELATED WORK

Predicting the economic life cycle has been an area of intensive research. Prior work discusses several financial time series models such as ARCH models, GARCH models [1], etc. all of which aim to identify subliminal rules governing markets. However relying on just financial time series data may not be sufficient as there are several other factors that might contribute to market fluctuations. Bollen et al. [2] argue that there is a strong correlation between stock market and public sentiment collected from Twitter. However, since location information is unavailable for around 40% of Twitter users [3], it is difficult to target predictions for a particular country. Preis et al. [4] on the other hand aim to find a link between weekly search engine query data and financial market fluctuations. They use keywords specific to particular companies to make inferences but this is not a good option for making inferences at the country level.

Lavrenko et al. [5] associate news with various stock trends and predict stock fluctuations by matching news patterns with trend labels. However, simply aligning news with various trends is too crude to predict specific topics, especially with respect to currency prediction systems which rely on several topic movements. Like Lavrenko et al., we believe that financial markets are sensitive to economic news but are also influenced by a wide variety of unanticipated events. While the market can react strongly to some news, it could remain completely insulated from other financial news. The aim of the FF system is to explore correlative links between news and financial market fluctuations. In particular it identifies cardinal topics from news articles which might have big

[1]Demo Address: `http://embers.cs.vt.edu/embers/alerts/visualizer_fin`

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
ACM 978-1-4503-2174-7/13/08 ...$15.00.

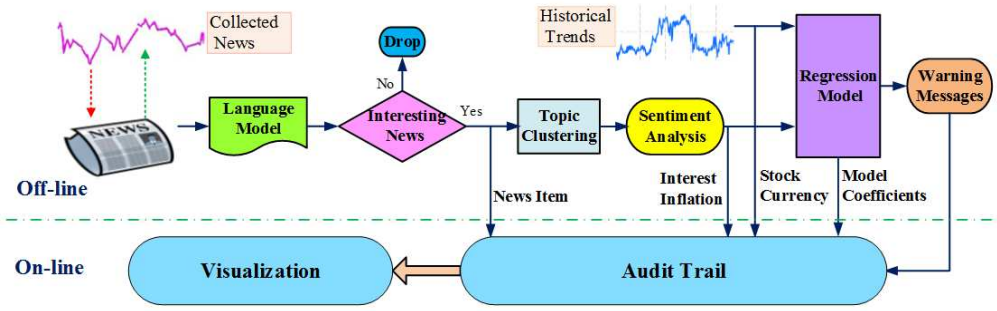

Figure 1: System framework, including off-line analysis and on-line audit trail generation.

impacts on markets and forecasts future market movements based on this information. The system primarily focuses on capturing significant fluctuations whose absolute value of $Zscore(30)$ (i.e., the number of standard deviation movements over the past 30 days) is greater than three in currency exchange rates which we refer to as OSI Events.

3. FRAMEWORK AND METHODS

As shown in Figure 1, the architecture of the Forex-foreteller system can be divided into two parts: off-line analysis and an on-line audit trail. The off-line analysis stage monitors news sources and employs language models to generate a pool of relevant news articles. Then we apply topic clustering methods and use customized sentiment dictionaries to uncover sentiment trends by analyzing relevant sentences. A linear regression model estimates the weight for each topic and makes currency forecasts. Artifacts from each stage including related news, topics, historical data, and model coefficients sent to the on-line stage along with the final warning for analysis.

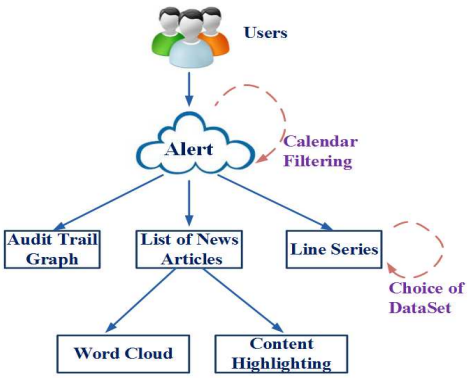

Figure 2: Audit trail workflow.

During the on-line stage, as shown in Figure 2, the FF system presents a list of generated alerts which, on selection, displays a dynamically generated graph symbolizing the inputs and outputs of each stage of the warning generation along with views to the actual raw data and news content. With the on-line stage, FF allows a user to analyze the historical financial and news data upon which an alert was founded.

3.1 Language Model

Not all financial news articles will influence a particular currency market (equally) and hence our first goal is to distinguish news articles that can impact currency movements. We develop a language model that classifies the incoming news articles as relevant or not relevant. We collect 361,782 Bloomberg news articles between April 2010 and March 2013. Using the latent Dirichlet Allocation (LDA) model of Blei et al. [6], we classify these news articles into 30 topics and obtain each article's topic distribution. Then, we identify top topics by manually aligning news articles with currency fluctuations, which are labeled as relevant topics. In order to classify incoming news articles, we estimate the topic distribution of each article and then decide whether its most prevalent topics fall into the set of relevant topics identified earlier.

3.2 Topic Tracking

The FF system tries to identify events that may influence the currency market. As explained above, using the LDA model, we obtain each document's topic distribution. Because we have reduced a corpus of news articles into a set of topics, it is straightforward to analyze the dynamics of those topics as a means of gaining insight into general topic distribution (e.g., see [7]). By tracking the topic distribution movement in time series and comparing with the mean topic distribution value, we identify emerging topics.

3.3 Sentiment Analysis

Interest rates, inflation values, and unanticipated events are all macroscopic indicators which are often reported in the news, either explicitly or implicitly. FF uses a novel approach to measure topic movements and fluctuations using sentiment analysis. The system identifies relevant sentences from news articles by filtering through a keyword list and then uses customized sentiment dictionaries to calculate movement (delta) values for interest and inflation rates. In this system, we use the Loughran-McDonald financial dictionary [8] to identify relevant keywords about interest rates and inflation values. For identifying unanticipated events, we use the AFINN dictionary [9] which is commonly used to measure large scale general emotions.

3.4 Forecasting

In FF system, we use a simple linear regression model to model the relationship between topic movements and currency fluctuations. (More complex models could be substituted here.) Our explanatory variables are interest rate,

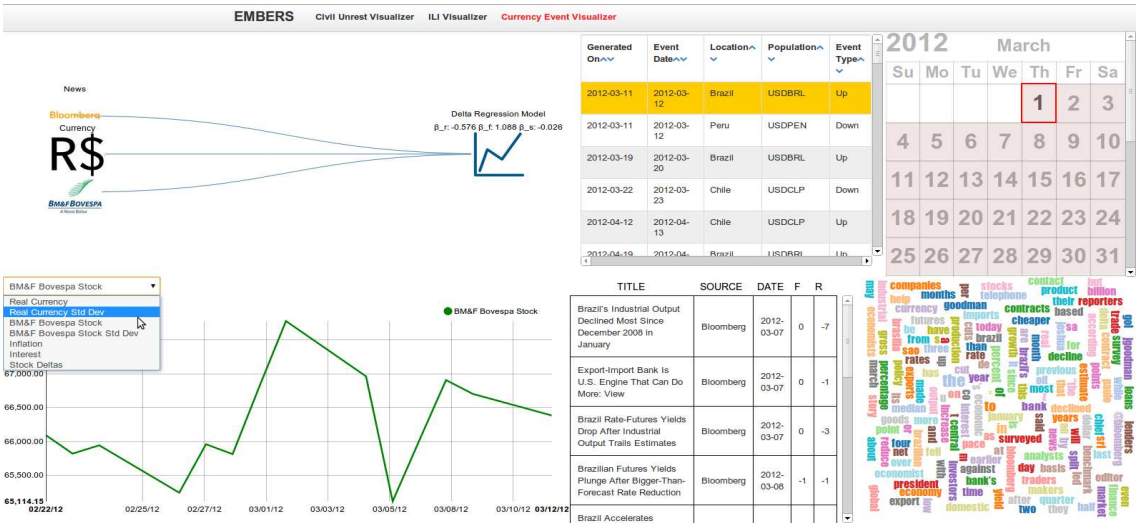

Figure 3: Forex-foreteller system snapshot.

inflation, stock and unanticipated events and the dependent variable is the change in currency value. We denote currency change as Δ_c, interest rate change as Δ_r, inflation change as Δ_f, stock market value change as Δ_s, and unanticipated events as Δ_e. Their respective weights are $\beta_r, \beta_f, \beta_s, \beta_e$. We use the the previous day's values (previous working day when the currency markets are trading) to forecast a given day's currency movement. Our linear function can be expressed as:

$$\Delta_c(t+1) = \beta_r \Delta_r(t) + \beta_f \Delta_f(t) + \beta_s \Delta_{\log s(t)} + \beta_e \Delta_e(t)$$

We always use the past two weeks historical data to fit the linear model, so each weight can be updated over time to keep up with a changing market.

3.5 Audit Trails

An important aspect of our system is the preservation and visualization of audit trails, shown in Figure 3. The system preserves all processed data and also provides a visualization. Figure 4 shows the 3-tier database structure for storing audit trails. Raw inputs store article level sentiment scores. Surrogates store stock values along with accumulated per-day sentiment scores for events, interest rates, and inflation values. Scores from all the lower tiers are used to generate warnings in the top tier. This structure gives us the flexibility to trace information reduction all the way back to the raw information.

3.6 Implementation Details

In the off-line FF stage, we have built a flexible, scalable pipeline of components that processes and manages raw data and computes anticipated currency fluctuations, as shown in Figure 5. There are four major component types: standard components, model specific components, queues, and databases. Standard components can be reused by other models and include news collection; enrichment for finding sentence boundaries, word lemmatization, and stemming; a key phrase matcher for matching keywords against our list, and locating their positions in a document; and the language model for classifying news. Queues are used to exchange processed data between different modules. All the components form a pipeline that begins with the collection of news articles and ends with anticipated currency fluctuations.

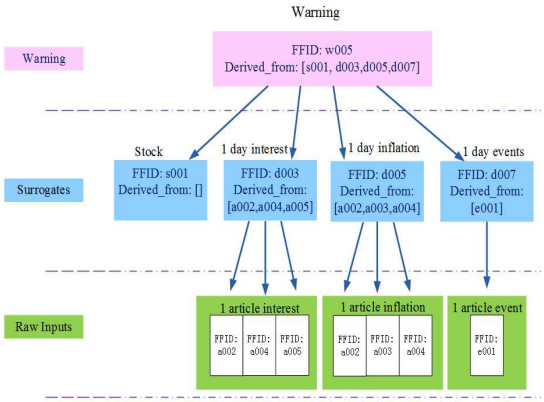

Figure 4: Database layers in FF.

4. DEMONSTRATION

4.1 Experimental Results

There is as expected a trade-off between recall and precision in our experiments. In FF, because we are trying to forecasts future events, we are more interested in high recall over precision. Using this system, we made predictions for Argentina, Brazil, Chile and Columbia from January 1, 2012, through December 31, 2012. The American dollar was used as the baseline currency. Summarized prediction results are shown in Table 1. For example, Columbia had two real three-sigma events and FF generated eight warnings, thereby leading to a low precision of 0.25. However, both the real events were correctly forecast leading to a high recall of 1.

4.2 Visualization

A web-based visualization (Figure 3) was developed to present market prediction outcomes, as well as the economic

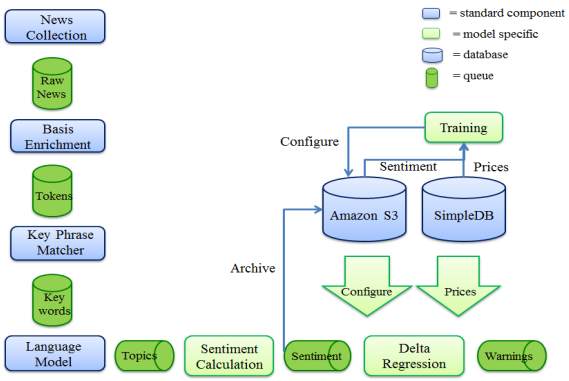

Figure 5: Off-line components and data flow.

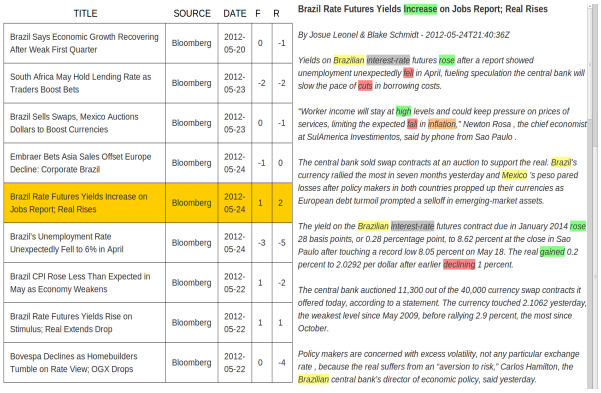

Figure 6: Inspecting news sources contributing to a warning. Content is highlighted based on each term's contribution: positive and negative terms are green and red, respectively; interest terms are gray, inflation terms are orange, and location indicators are yellow. "F" for delta inflation, "R" for delta interest rate.

Table 1: Experimental Results

Country	Warnings	Events	Matches	Precision	Recall
Argentina	17	5	3	0.18	0.60
Brazil	18	8	5	0.28	0.63
Chile	3	1	1	0.33	1
Colombia	8	2	2	0.25	1

background including historical stock and currency valuations, and relevant news articles. The visualization is comprised of a context area on the upper half and a details area below. Users can explore predicted *OSI* events and minimize search scope by setting lower and upper bounds with the calender. Once the user has selected an *OSI* event, a diagram of the inputs and outputs that make up the audit trail will be generated on the left. Historical economic trends that factored into the *OSI* event, including the original currency and stock values, the standard deviations for each, and values for inflation and interest changes computed from news items are displayed on the left of the details area. To the right the interface shows the news articles relevant to the forecast. Initially, a word cloud of frequent unigrams from all relevant news items is displayed; when a user chooses an article from the list, the right corner will show the content of this news item along with highlighted words that contributed to inflation, interest, and other factors.

4.3 Use Cases

Forex-foreteller was able to forecast most of the studied appreciations and depreciations. For instance, the year 2012 saw the Brazilian Real's (BRL) exchange rate significantly altered due to government interventions. On May 21st, the system predicted that the BRL would continue depreciating correctly as per the trends from the previous couple of weeks. However, on May 22nd, as the Brazilian government was intervening to reverse the Real's fall, FF was able to correctly forecast the reversal of the currency movement and predicted that the currency would appreciate. On May 24th it correctly forecast that the currency would appreciate on May 25th. Figures 6 shows the news articles used to make the prediction for May 25th in order to give an idea of how the warning was generated.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

5. REFERENCES

[1] R. S. Tsay, *Analysis of financial time series*. Wiley-Interscience, 2010, vol. 543.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[3] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proc. CIKM'10*. ACM, 2010, pp. 759–768.

[4] T. Preis, D. Reith, and H. E. Stanley, "Complex dynamics of our economic life on different scales: insights from search engine query data," *Phil Trans Math Phys Eng Sci*, vol. 368, no. 1933, pp. 5707–5719, 2010.

[5] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, "Mining of concurrent text and time series," in *KDD'00 Workshop*. Citeseer, 2000, pp. 37–44.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[7] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. PNAS*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[8] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[9] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *ARXIV*, vol. 1103, no. 2903, 2011.