Exchange Rate Modelling Using News Articles and Economic Data

Debbie Zhang, Simeon J. Simoff, and John Debenham

Faculty of Information Technology, University of Technology, Sydney
{debbiez, simeon, debenham}@it.uts.edu.au

Abstract. This paper provides a framework of using news articles and economic data to model the exchange rate changes between Euro and US dollars. Many studies have conducted on the approach of regressing exchange rate movement using numerical data such as macroeconomic indicators. However, this approach is effective in studying the long term trend of the movement but not so accurate in short to middle term behaviour. Recent research suggests that the market daily movement is the result of the market reaction to the daily news. In this paper, it is proposed to use text mining methods to incorporate the daily economic news as well as economic and political events into the prediction model. While this type of news is not included in most of existing models due to its non-quantitative nature, it has important influence in short to middle terms of market behaviour. It is expected that this approach will lead to an exchange rate model with improved accuracy.

1 Introduction

Exchange rates prediction is one of the most challenging applications of modern time series forecasting. Despite a large amount of research being done in the area, economists know remarkably little about exchange rate regimes [1]. Meese and Rogoff [2] in their well-known paper showed that fundamentals (economic index) dictated by monetary models do not outperform a naive model of no changes in the out-of-sample forecast of nominal exchange rates. After over twenty years of research since the publication of the Meese-Rogoff studies, their findings remain very robust. Although the structural models do not deliver a better forecasting performance than a random walk model, there is evidence on the ability of the structural models to correctly predict the direction of change. Among the enormous amount of empirical models, the sticky price monetary model of Dornbusch and Frankel remains the workhorse of policy-oriented analysis of exchange rate fluctuations [3], which can be expressed as follows:

$$s_t = \beta_0 + \beta_1 \hat{m}_t + \beta_2 \hat{y}_t + \beta_3 \hat{i}_t + \beta_4 \hat{\pi}_t + \mu_t \tag{1}$$

where s is the changes of interest rate during each sampling period; m and y denote the logarithm of the money value and real GDP respectively; i and π are the interest and inflation rate, respectively; $\hat{\bullet}$ denotes the inter-country difference of the corresponding variable; μ is the error term.

S. Zhang and R. Jarvis (Eds.): AI 2005, LNAI 3809, pp. 467–476, 2005.

While models based on fundamentals have performed reasonably well in explaining exchange rate development in the long term, economists have little success in predicting exchange rate in short and middle term movement. The general consensus of the poor performance of the traditional empirical models using economic fundamentals to account for exchange rate developments on short to medium term is caused by the irrationality of the market participants, bubbles, and herd behaviour, which are hard to be captured in econometric models.

Recent literature shows that news about fundamentals has played an important role in creating market dynamics. Prast and De Vor [4] have studied the reaction of investors in foreign exchange markets to news information about the euro area and the United States on days of large changes in the euro-dollar exchange rates. Unlike the traditional models, daily changes in the euro/dollar rate on news about economic variables in the United States and the euro area, and the variables capturing news in the two economies were used in the regression model, which is:

$$E_t = \alpha + \sum_{i=1}^{8} \beta_i D_i + \varepsilon \tag{2}$$

where E_t is the percentage daily change in the euro-dollar exchange rate; D_{1-8} represent the following variables: 1 - real economy, euro area; 2 - inflation, euro area; 3 - change in official interest rate, ECB; 4 - statements/political events, euro area; 5 - real economy, United States; 6 - inflation, United States; 7 - change in official interest rate, United States; 8 - statements/political events, United States. It has been found that there is strong correlation between exchange rate daily movement and the market participants' responses to the daily economy news and political events.

More recent research has confirmed that news has statistically significant effects on daily exchange rate movement. Ehrmann and Fratzscher [5] have evaluated the overall impact of macro news by analysing the daily exchange rate responses using similar regression models with news variables. Three key results were found. Firstly, the news about fundamentals can explain relatively well the direction, but only a much smaller extent to the magnitude of exchange rate development. Secondly, news about US economy has a larger impact on exchange rates than news about the euro area. Thirdly, higher degree of market uncertainty will lead to more significant effects of news releases on exchange rate movements.

The above findings motivated the research reported in this paper. By using the text mining techniques, the manual process of identification and classification of positive and negative news can be automated. As the correlation between news and currency exchange rate has only been identified recently, there is not much work reported in this area. Eddelbüttel [6] and Wong [7] both tried to use the keywords in news headlines to forecast intraday currency exchange rate movements. Eddelbüttel used a set of keywords to identify the relevant news and sorted them into three groups: "All", "DEM" and "USD". Then the number of news pieces in three groups are calculated and used as the variable in

the GARCH(1,1) model for prediction. The news analysis is restricted to the counting of the number of relevant news headlines to avoid qualitative judgement about "good" and "bad" news. Wong etc. proposed a prediction model based on the occurrence of keywords in news titles. The keywords in the news title are identified by selecting the words with the highest weighting values. A set of rules for predicting the exchange rate movement direction from the keywords in the news titles are generated. These over simplified approaches only utilise news information to a very limited extent. In this paper, a more sophisticated text mining approach for news filtering and classification is presented. An empirical model based on macroeconomic data and the results of new classification is proposed. The system structure and implementation issues are also provided.

2 News Filtering and Classification Based on Text Mining

Before incorporating the news effect into an exchange rate model, it is important to identify the relevant news and classify them into "good" or "bad" news category, that would have opposite impact on the market behaviours. This section describes the training process of news filtering and classification.

2.1 Data Collection and Pre-processing

News articles used in the prediction model are retrieved from online news sources. Prior to the processing, the news articles used for training are manually classified into two groups: news affect exchange rate (target corpus) and other news (generic corpus). Choosing the news articles in the target corpus is crucial for the process since the target corpus contains the underlying knowledge of what factors affect exchange rate movement. Much research has studied the factors that affect currency exchange rate, which can be macroeconomic data, statements by central bankers and politicians and political events that affect macroeconomics. Therefore, only the news that is relevant to these is chosen.

To improve the process efficiency and avoid noise distraction, stop words in the target corpora are replaced by a stop word symbol but are not removed completely to avoid incorrect word co-occurrence. Porter stem algorithm is also applied to remove the common morphological and inflexional endings from words in the documents.

2.2 Automatical Keyword Extraction

Text mining operations are mainly based on the frequency of keywords. The goal of this step is to generate the best set of keywords that can distinguish news documents related to exchange rate from other news documents. To reduce the calculation complexity and increase the processing efficiency, the number of keywords are kept to the minimum amount but are still a good approximation of the original document set in its full space. There are two types of keyword

frequencies used in this paper: term frequency and document frequency. The term frequency is calculated by the number of times a term appears in the corpora. The document frequency is the number of the documents that contain this term in the corpora.

Keywords are not restricted to single words, but can be phrases. Therefore, the first step is to identify phrases in the target corpus. The phrases are extracted based on the assumption that two constituent words form a collocation if they co-occur a lot [8].

Once the phrases have been extracted, the key terms are extracted amongst the single words except stop words and phrases in the target corpus. The generic corpus as the background filter. The distribution of terms in the target corpus and the generic corpus are compared. The terms in the target corpus that stand out are considered as the features of the corpus, indicating that these terms are domain-specific terminology.

Dunning [9] suggested the log likelihood ratio (LLR) Chi-square statistic test is effective in determination of domain-specific terms. Vogel [10] also reported that LLR had a greater ability to differentiate the importance of a term in a domain than other methods such as information gain (IG) or mutual information (MI).

The likelihood ratio for a hypothesis is the ratio of the maximum value of the likelihood function over the subspace represented by the hypothesis to the maximum value of the likelihood function over the entire parameter space. In this case, the null hypothesis H_0 is formulated to test the distribution of a term is the same in the generic corpus and target corpus. H_a measures the actual distribution of the term in the whole data set. The log likelihood ratio for this test is:

$$-2\log\left(\frac{H_0\left(p;k_1,n_1,k_2,n_2\right)}{H_a\left(p_1,p_2;k_1,n_1,k_2,n_2\right)}\right) \tag{3}$$

The binomial distribution of the log likelihood statistic is given by:

$$\begin{aligned}-2\log\lambda \;&=\; 2\left[\log L\left(p_1,k_1,n_1\right)+\log L\left(p_2,k_2,n_2\right)\right.\\ &\quad\left.-\log L\left(p,k_1,n_1\right)-\log L\left(p,k_2,n_2\right)\right]\end{aligned} \tag{4}$$

where $\log L(p,n,k) = k\log p + (n-k)\log(1-p)$,
k_1 and k_2 are the document frequency of a term in the target corpus and generic corpus respectively,
n_1 and n_2 are the size of the target corpus and generic corpus respectively,
$p_1 = \frac{k_1}{n_1}$, $p_2 = \frac{k_2}{n_2}$, and $p = \frac{k_1+k_2}{n_1+n_2}$.

The method scores the terms based on the difference in the percentage of documents containing the term in the target and generic corpus. It does not distinguish whether the difference is caused by the term occurring more or less in the target corpus. As in this research that only the terms significant in the

target corpus are concerned, a simple condition is added to the ranking equation so the terms are significant in the generic corpus are filtered out:

$$\frac{p_1}{p_2} \geq 1 \tag{5}$$

2.3 News Relevance Classification

The news relevance classification is divided into two steps: the first step is to identify the news that has potential to cause movement in exchange rates, the second step is to identify the news that is Euro and/or US dollar related.

The exchange rate related news can be separated from other news based on the key terms extracted from the previous section, which often well represent the characteristics of the data set. In this case, a modified k-Means classification algorithm, which is particular suitable for this case, is chosen as being computationally simple and efficient. The centroid of the target corpus and the maximum Euclidean distance in the training data are calculated. The maximum distance is used as the threshold to determine if the data belongs to a target cluster.

News related to exchange rate may not be discussing Euro and US dollar currencies, which is further identified by using the frequency of the words of currency and country names it contains.

2.4 Positive and Negative News Classification

It is important to further classify the relevant news into "positive news" and "negative news" categories, as news in different groups have entirely different effects on the market behaviour.

Recent studies show that the effect of the news is the combined effect of market expectation and the news itself. A piece of news could have positive or negative impact to the market depending on the market expectation. Therefore, unlike some studies that define good and bad news by its immediately effect to the market, in this research, the news is defined to be good or bad according to its fundamental effect to the market. The market expectation is incorporated into the model in a later stage. For example, a news about US increased its interest rate is defined to be positive news to US dollars.

The task of identifying "good" and "bad" news of exchange rate is not straight forward since both groups of news use similar set of keywords. For example, the following two pieces of news have exactly same set of words, but one is considered to be positive and the other is considered to be negative to the appreciation of US dollars:

1. The interest rate has gone up. The US dollar has gone down.
2. The interest rate has gone down. The US dollar has gone up.

The positive and negative news can use similar set of key terms, which causes great difficulties in the classification. However, the sequences of the key terms

can represent the meaning of sentences better, which is well illustrated in the previous example. Therefore, a term is defined as the sequence of key terms in a sentence, which is used as the input features for the positive and negative news classification. The feature vectors of the above example can be represented as:

Table 1. Example of feature vector representation for "good"/"bad" news classification

features	document 1	document 2
interest rate up	1	0
interest rate down	0	1
US dollar down	1	0
US dollar up	0	1

However, using key term sequence as classification features leads to a high dimensional vector space with sparsely distributed elements, which causes difficulty in separating instances into classes (subspaces). Therefore, the discriminant analysis is implemented to combine features of the original data in a way that most effectively discriminates between classes [11]. The discriminant analysis is to project the documents onto a lower dimensional subspace of the original vector space. After the projection, instances in the same class are tightly grouped, but well separated from the other clusters. Also, with a smaller set of input features, the complexity of the classification is reduced and the calculation efficiency can be greatly improved.

Again, the document collection with n documents and m features in cluster i represented by a term (key term sequence) frequency document matrix. In this application, there are two clusters -"good" news and "bad" news.

$$A_i = [a_1 a_2 \cdots a_{n_i}] \in \Re^{m \times n_i} \tag{6}$$

An optimal linear transformation G^T can be found such that the Euclidean distance between the clusters is maximised while the distance between instances within each cluster is minimised:

$$G^T \in \Re^{l \times m} : a_j \in \Re^{m \times 1} \rightarrow y_j \in \Re^{l \times 1}, 1 \leq j \leq n_1 + n_2 \tag{7}$$

To measure the cluster quality, scatter matrices are formulated based on the distance of each instance to the centroid. The scatter matrix within cluster and between clusters are defined as the following equations:

$$S_w = \sum_{i=1}^{2} \sum_{j \in n_i} \left(a_j - c^i\right)\left(a_j - c^i\right)^T \tag{8}$$

$$S_b = \sum_{i=1}^{2} \sum_{j \in n_i} \left(c^i - c\right)\left(c^i - c\right)^T = \sum_{i=1}^{2} n_i \left(c^i - c\right)\left(c^i - c\right)^T \tag{9}$$

The scatter matrices of the transformed feature vectors are as follows:

$$S_w^Y = G^T S_w G, \quad S_b^Y = G^T S_b G \tag{10}$$

The closeness of the instances with the cluster and the separation between clusters can be calculated from the scatter matrices as trace(S_w^Y) and trace(S_b^Y) respectively.

The transformation matrix G^T is calculated by maximising the value of trace $S_w^Y \left(S_b^Y\right)^{-1}$ that approximates the maximisation of $trace(S_w^Y)$ and minimisation of $trace(S_b^Y)$. The numerical algorithm for this optimisation problem presented in [11] is adapted.

After the G^T being calculated, the k-Means classification algorithm can be applied to classify the transformed feature vectors y_j into "good" and "bad" news categories.

3 The Econometric Model of Exchange Rate Responding to News and Economic Data

This research focuses on using text mining methods to incorporate the information in the news articles into a currency prediction model. As euro/dollar exchange rate will be used as the testing case, the empirical model presented by Galati and Ho to study the news effect on economic data particular for euro/dollar exchange rate is chosen [12]. In this work, the above model is modified to incorporate a news index (I_{news}), which reflects the news effect on exchange rate. The regression equation has the following form:

$$\Delta \ln S(t) = \alpha_0 + \alpha_i x_i(t) + \beta I_{news} + \varepsilon \tag{11}$$

where x_i represent the economic data variables which include: US non-farm payrolls, US unemployment rate, US employment cost index, US durable goods orders, NAPM manufacturing, NAPM non-manufacturing, US advance retail sales, US industrial production, US CPI, Ifo index, Germany unemployment rate, Germany industrial production, INSEE industrial trends, Germany CPI and EU 11 PPI.

4 System Structure and Implementation Issues

The system is designed as a multi-agent system, as shown in figure 1.

The user interface module is the accessing point to the system, which is shown in Figure 2. The exchange rate data and the news articles are displayed on the main page. There are three major menu items: "Data", "Training" and "Prediction Model" on the menu bar. Each of these menu items controls one of the agents in the system.

"Data" menu item allows the user to set the data sources (URLs) and schedule the download frequency and time for the data extraction agent. Since users

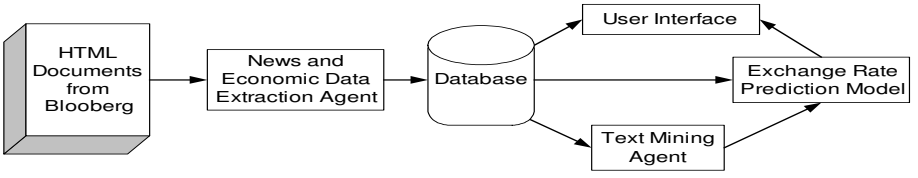

Fig. 1. The structure of the exchange rate prediction system

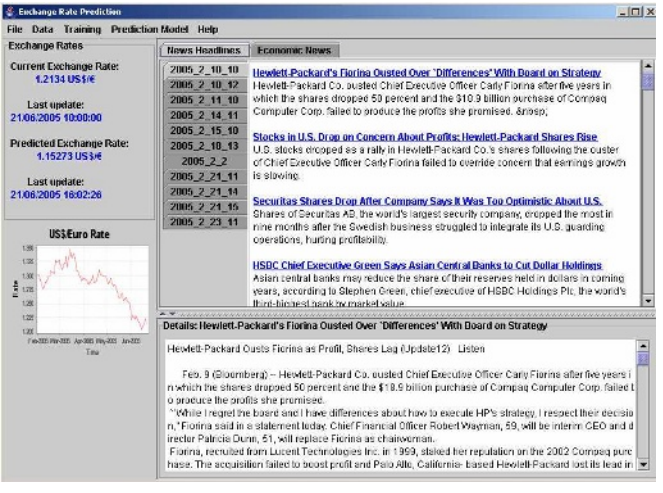

Fig. 2. User interface

are allowed to set the data sources, the template of the HTML files retrieved from the data source is unknown. Therefore, flexible methods to extract the data from HTML tags are required. Extracting the news articles from HTML file is not a simple task since each web site has different layout and format. A generic news extraction and validation algorithm was developed to solve this problem [13]. Extracting the economic data from HTML files is relative simple since data are always held in HTML tables. A program was developed to search these tables by using keywords.

"Training" menu item is designed to execute the operations provided by the text mining agent. It contains a pull-down menu list with five options: "Select Good/Bad News", "Keyword Generation", "News Relevancy Training", "Good/Bad News Training" and "Automatic Training". Each of the menu item executes one of the training step described in section 2.

Once the training process is completed, the above classification system can be used to classify "live" news articles online, as shown in the following algorithm:

```
Input: news article tt (String)

Output: news category n (integer 0<=n<=2)
        when n=0, unrelated news
        when n=1, positive news
        when n=2, negative news
    const
        currency keywords US_Euro (String[])
        currency keywords Non_US_Euro (String[])
    var
        generated from "Keyword Generation":
            key terms kw (String[])
        generated from "News Relevancy Training":
            centroid C (float[])
            threshold T (float)
        generated from "Good/Bad News Training":
            key term sequences ks (String[])
            transformation matrix G_T (float[][])
            "good" news centroid g_C (float[])
            "bad" news centroid b_C (float[])

Begin:  apply stem algorithm to tt
        replace all words in tt to a word symbol except words in kw
        calculate the frequency of the key terms kw in tt
        d := the distance to the relevance news data set centroid C
        if d > T
            n = 0 (unrelated news), return;
        calculate the frequency of currency keywords US_Euro
        calculate the frequency of currency keywords Non_US_Euro
        compare the average frequency of US_Euro and Non_US_Euro
        if average frequency Non_US_Euro > average frequency US_Euro
            n = 0 (unrelated news), return;
        remove all quotations in tt except """, ",", "?", "."
        A := calculated key term sequence ks frequency in tt
        Y := G_T * A
        d_good = distance of Y to "good" news data set centroid g_C
        d_bad = distance of Y to "bad" news data set centroid b_C
        if d_good > d_bad
            n = 2 (negative news), return;
        else
            n = 1 (positive news), return;
End.
```

The exchange rate prediction model is generated offline by a Matlab program. The "Prediction model" menu item allows users to input the regression parameters of the model. At the end of each day, a prediction value is calculated based on the number of "good"/"bad" news documents of the day and the current economic data.

5 Conclusions

A novel approach to an exchange rate prediction model using news articles and economical data has been developed. This paper focuses on the aspects of theoretical development and system structure design. A comprehensive case study is being conducted to evaluate the system.

References

1. Rose, A.K.: Exchange rate regimes and stability: Where do we stand? Technical Report Unpublished working report, U.C. Berkeley. (2004)
2. Meese, R., Rogoff, K.: Empirical exchange rate models of the seventies. do they fit out of sample? Journal of International Economics **14** (1983) 3–24
3. Dornbusch, R.: Expectations and exchange rate dynamics. Journal of Political Economy **84** (1976) 1161–1176
4. Prast, H.M., de Vor, M.P.H.: Investor reactions to news: a cognitive dissonance analysis of the euro-dollar exchange rate. European Journal of Political Economy **21** (2005) 115–141 TY - JOUR.
5. Ehrmann, M., Fratzscher, M.: Exchange rates and fundamentals: new evidence from real-time data. Journal of International Money and Finance **24** (2005) 317–341 TY - JOUR.
6. Eddelbttel, D., McCurdy, T.: The impact of news on foreign exchange rates: evidence from high frequency data. Technical report, University of Toronto (1998)
7. Peramunetilleke, D., Wong, R.K.: Currency exchange rate forecasting from news headlines. Aust. Comput. Sci. Commun. **24** (2002) 131–139
8. Manning, C.D., Schutze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, Mass. (1999) Christopher D. Manning, Hinrich Schutze. 24 cm.
9. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics **19** (1994) 61–74
10. Vogel, D.: Using generic corpora to learn domain-specific terminology. In: The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Workshop on Link Analysis for Detecting Complex Behavior), Washington, DC, USA (2003)
11. Berry, M.W.: Survey of text mining : clustering, classification, and retrieval. Springer, New York (2003)
12. Galati, G., Ho, C.: Macroeconomic news and the euro/dollar exchange rate. Technical Report 105, Bank for International Settlements (2001)
13. Zhang, D., Simoff, S.: Informing the curious negotiator: Automatic news extraction from the internet. In: Australasian Data Mining Conference, Cairns, Australia (2004)