Forecasting Currency Exchange Rate Variations with News Articles using LSTM Model (Progress Report)

Nianmin Guo
Wanhui Han
{guo.ni,han.wa}@husky.neu.edu
Northeastern University

1 Changes

We have changed the main focus and title of our project. Upon further inspecting the expected input and output of our model, it seems like that designing a model that can produce the exchange rate variations is more achievable compared to analyzing the sentiment of the articles first. Sentiment analysis, in this case, will bring a layer of abstraction to the model, making it potentially difficult for the downstream model to generate an accurate exchange rate prediction. In addition, since there is no ground truth for this prediction, it might not be feasible to generate an accurate sentiment analysis model. Thus, we have decided to narrow our focus to News to Exchange Rate variation.

2 Data Preprocessing

`R` is used for data preprocessing for both News and Exchange Rate datasets. The time-sensitive nature of exchange rates required us to remove all entries of which date information is unavailable. Year 2016, for example, yielded 223 valid entries for Pound to US Dollar exchange rate and 81,624 valid news articles. Then, each dataset underwent the following procedures:

2.1 News Articles

1. Removed irrelevant fields, such as *Publisher*, *Title*, and *URL*.
2. Clean up article content, including line breaks `\n`,`\r`, etc.
3. For the purpose of testing models, a subset of the data is used (Calendar year 2016).
4. The content of articles are then passed through a sentence tokenizer in `R` [3], with punctuation and capitalization removed.
5. Any sentence shorter than 30 characters is removed, as the tokenizer package in `R` sometimes erroneously ends sentences on abbreviation marks .
6. With the steps above, the dataset is comprised of 2 columns – *Date* (for joining), and *Content* as sentences.

2.2 Exchange Rates

1. Filter the major currencies (For the final project we are looking to include all major currencies such as Pound, Euro, Japanese Yen, Chinese Yuan, and Canadian Dollar). For the purpose of testing code structure, Pound is used.
2. Generate `day_before` tag, which compares the percent change of one day's rate to the day before; and returns a string tag with the following rule:

Change	Tag	%
<= -2.5%	NL	1.5
(-2.5%, -1.0%]	NM	4.87
(-1.0%, -0.5%]	NS	15.85
(-0.5%, 0%]	NX	31.74
[0%, 0.5%)	PX	25.63
[0.5%, 1%)	PS	13.65
[1%, 2.5%)	PM	6.38
>= 2.5%	PL	0.37

2.3 Combining Datasets

1. The two datasets are joined by date
2. With the steps above, the dataset is comprised of sentences and their corresponding exchange rate variation tags.

3 Methods

We have determined to use a LSTM model that takes article sentences and date as input and tags as output. We are using the LSTM model by Tensorflow in the first steps with 64 hidden states with the GloVe vectors with 400k vocab [5]. We are choosing LSTM for its ability to capture the sentiment in long sentences [6], as well as its ability to handle time series well [4]. We think date/time is essential since news from one day to the next often relate to each other, but we are looking for ways that allow us to combine both pieces together as input.

4 Results

As of right now, we are able to train the model with articles and their corresponding tags. However, with 223 exchange rate data points and over 80,000 articles, the model is ineffective for predicting exchange rates variations (40% precision). It is possible that the input data, at sentence level, is too noisy for useful predictions; or it might be due to the lack of resolution in exchange rates data.

5 Improvements

5.1 Model

We will be continuing to use LSTM model, with changes made to the input data first. However, with multi-currency prediction we may be required to make further changes to the model so it can produce multiple tags at once.

5.2 News Data

In our next iteration, we will be exploring with article-level content instead of sentence-level content, since it is hard to derive market variation information from most individual sentences. Since our current dataset included 13 news sources including *Verge* and *Breitbart*, of which are either less related to market or heavily biased, We are also filtering the articles by outlet.

5.3 Exchange Rates

We are looking to obtain historical exchange rates of higher time resolution. However, coordination with the news data collected is also needed so that time resolution of both datasets can be matched.

6 Future Steps

We will be working on creating a new dataset first with article level input and higher relevance. With a higher-than-chance precision we will then tweak the neural network by using a larger vector model (1.9M GloVe or Google word2vec). With promising results we will be working on training the model with data of multiple major currencies.

The stretch goal of predicting the effective terms (i.e., interval of one week instead of one day) of a news article will still be pursued.

References

[1] Junfeng Jiang and Jiahao Li. 2018. Constructing Financial Sentimental Factors in Chinese Market Using Natural Language Processing. (2018).

[2] Huicheng Liu. 2018. Leveraging Financial News for Stock Trend Prediction with Attention-Based Recurrent Neural Network. (2018).

[3] Lincoln A. Mullen, Kenneth Benoit, Os Keyes, Dmitry Selivanov, and Jeffrey Arnold. 2018. Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software* 3 (2018), 655. Issue 23. https://doi.org/10.21105/joss.00655

[4] Lina Ni, Yujie Li, Xiao Wang, Jinquan Zhang, Jiguo Yu, and Chengming Qi. 2019. Forecasting of Forex Time Series Data Based on Deep Learning. *Procedia Computer Science* 147 (2019), 647–652. https://doi.org/10.1016/j.procs.2019.01.189

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[6] Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. [n. d.]. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 225–230.

[7] Yequan Wang, Minlie Huang, and Li Zhao. [n. d.]. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 606–615.