Quantification of Financial News for Economic Surveys

Mihail Minev
Interdisciplinary Lab for
Intelligent and Adaptive Systems
Computer Science and
Communications Research Unit
University of Luxembourg
Email: mihail.minev@uni.lu

***Abstract*—This study concerns financial news articles, which reflect the monetary policy during the US subprime mortgage crisis. In particular we consider official announcements conducted by the Federal Reserve and its leading representatives. We aim to quantify such information using dependency parsing techniques and statistical measures. In addition, we examine the correlations between the monetary policy and the stock markets by modeling composite index volatilities as functions of key publications. A prototype for the classification of news is targeted, which should reveal the economical impact of events. An eminent aspect of our study is the identification, extraction, and representation of topic-related features and the corresponding instances.**

***Keywords*-feature extraction, financial news, classification**

I. Introduction

Recently we observe a continuous increase of news publications. Information about significant events is spreading across the globe within milliseconds. However, news articles are written by humans and the content is explicitly addressed to human readers. As a result, a normal machine is incapable of interpreting and classifying these documents without a selective pre-processing. In order to enable surveys on the central bank's fiscal policy, we describe next a linguistically motivated approach for the equivalent announcements quantification.

In this paper we focus on the feature annotation of financial news related to the monetary policy of the United States. In particular, we address the period between 2007–2012, which captures the beginning and the development of the subprime mortgage crisis (alias financial crisis). In line with this events, we analyze announcements from the Federal Reserve—the central bank in the US. The main goal of this work is to quantify the principal facts in the news documents and thus to enable the execution of economic surveys within the monetary policy domain.

Due to the natural ambiguity of financial news, we adopt a combination of natural language processing techniques and statistical measures. First, we identify and extract multi-word terms (alias features) from the data, which we aggregate to a domain-specific terminology. And second, we complement these features by determining their conditional instances in each document. Finally, we train a multi-instance (MI) classifier, which correlates the fiscal policy decisions with abnormal stock market movements.

The paper is organized as follows: Section II provides a literature overview focusing on news quantification as conducted in cross-discipline studies—in particular computer science and finance; Section III outlines the data acquisition, the project architecture and the corresponding research workflow; Section IV summarizes the study achievements and outlines potential future works.

II. Related Works

In the related literature, the quantification of financial news is primarily dedicated to the prediction of stock market prices, trends and volatilities [1]. The developed prototypes attempt to mimic human reasoning by evaluating text articles and determining patterns in historic data. In this context, the extraction of definite features in financial news plays a key role. Another crucial factor influencing the prediction performance of the prototypes is the identification of correlations between those features and the tick data. Consequently, in this research overview we emphasize on the text representation methods and the news classification algorithms.

In an early study [2] analyze articles from *'The Wall Street Journal'* and *'The Financial Times'*. For the text representation the authors use a handcrafted dictionary of 423 multi-word terms, which is composed by financial experts. The features include word tuples and are further weighted with TFxIDF. According to the study results, the k-Nearest Neighbor algorithm outperformed the Neural Networks by a small margin. Likewise, [3] utilize a vocabulary of 400 manually selected sequences of two to five words. However, the authors examine exclusively news headlines, which usually do not follow linguistic rules. Here, TFxCDF surpasses TFxIDF and a third boolean method.

The more recent studies in news quantification consider automatic feature extraction in combination with a state-of-the-art ranking algorithm. For instance, [4] examine n-grams and two word combinations, syntactically labeled as noun phrases. In order to reduce the feature space dimensionality, the authors apply the Chi-Square algorithm. In total, the SVM classifier performs better than Neural Networks and Naïve Bayes. Besides Chi-Square, [5] examine also Information Gain and TFxIDF. The first technique achieves the highest score in conjunction with LibSVM. Additional tests are conducted with k-Nearest Neighbor, Neural Networks, and Naïve Bayes. In contrast, [6] represent an event as a triple of: (1) an economic actor (e.g. company), (2) a verb and/or an adjective, and (3) an object (e.g. profits). The authors construct rules with regular expressions, which capture the market reactions on positive/negative events. Only these events are learned by a self-training algorithm, which achieves its best score exclusively on the story headlines. In the same context a comparative study [7] between three news representation techniques was published in 2009. The candidates are Bag-of-Words, Named Entities, and Noun Phrases. The latter scored best—in two out of three prediction metrics—linked with Support Vector Regression.

In summary, none of the presented studies considered advanced linguistic aspects like shallow or dependency parsing for the text representation. We argue, that a thorough semantic and syntactic analysis of the word relationships in a document will enable a more accurate term annotation and thus an enhanced classification.

III. The Model

A. Data Acquisition

We collected our data from the official web page [8] of the Federal Reserve Bank. The collection includes 174 documents with 1.225.719 tokens. Four document types are inclosed in the set. Federal Open Market Committee (FOMC) statements, which are released regularly eight times per year. The corresponding minutes, which are published three weeks later. The summary of the policymakers' economic projections, which are issued four times per year on the same date as the FOMC minutes. And finally the Beige Book, which is announced before each regular meeting and which aggregates economic reports by Federal Reserve district (twelve in total) and sector.

B. The Architecture

The architecture in Fig. 1 illustrates the study milestones, which are separated in four main categories. These include: (1) the document retrieval; (2) the information extraction; (3) the classifier setup; and (4) the model evaluation. In the next section we address the objectives for each one of them.

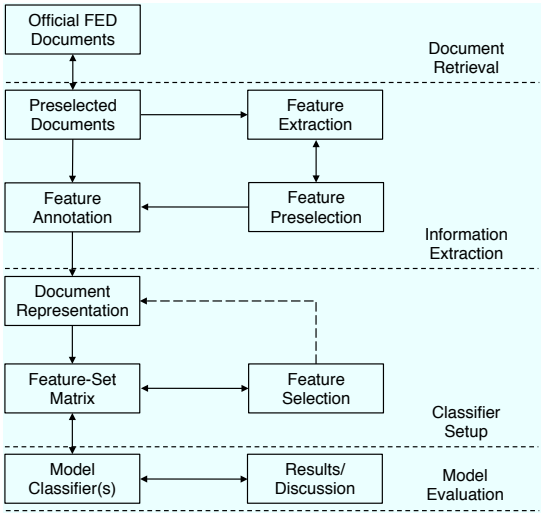

Figure 1. The research workflow includes the following major categories: 1) Document retrieval, where the four related document types are crawled; 2) Information Extraction, which includes the feature identification, preselection and annotation steps; 3) Classifier Setup, which describes the feature selection methods and the classifier configuration; 4) Model Evaluation, where we interpret and discuss the preliminary results from the study.

1) Document Retrieval: Numerous economic studies [9], [10], [11] indicate a link between the monetary policy decisions and asset prices. In order to identify precisely the triggers of such correlations and to enable further analysis with more independent variables, we collected four document classes from the Federal Reserve website. These include official monetary policy events such as the FOMC statements, the meeting minutes, the economic projections, and the Beige Book reports. In addition, for each document we record the publishing date and time.

2) Information Extraction: In this section we examine terminology extraction algorithms, which support linguistic and statistical aspects. First, we apply a part-of speech tagger and a shallow parser, which capture the semantic and syntactic term relationships. The result is a collection of grammatically annotated single- and multi-word phrases. We are primarily interested in extracting the noun phrases, which according to [12] are the most expressive language compounds and thus count as especially informative. Second, we weight the candidate features over the document set using the C-value [13] algorithm. For the term ranking, the method combines linguistic information with statistical measures as the frequency distribution and the word-length. Consequently we determine words and word combinations, which serve as a terminology for the monetary policy domain. In this particular case, the terminology consists of specific compounds, which cover principal indicators like 'unemployment rate', 'housing sector', 'federal funds rate', 'business spending', et cetera. We take advantage of this factor and acknowledge the domain-related terms as absolute variables, due to their neutral expressiveness.

After acquiring a set of features from the data, we conduct a search for the matching values. All sentences in a single document are represented as phrase structure grammar trees. Linguistic studies distinguish between four central phrase categories: noun, verb, adverb, and adjective. A phrase is defined by a coherent head word and is optionally extended by other words. The links between these phrases are extracted using a dependency parser trained on *WSJ*. As a result, for each identified variable we generate a bag of candidate values from the text. Hereby we further narrow the value selection by separately examining each subtree. For example, the feature 'housing sector' receives in one document the value 'is depressed', where in another it gets the contrasting value 'remained stable'. Consequently, we aggregate the feature-value pairs over the entire collection preserving also the attributes as the document type and its timestamp. In the last step we remove all stop words, which are non-deterministic.

Since not all candidate features meet the domain-specific requirements for economic indicators, it is unfeasible to appoint meaningful values to each of them. In order to reduce the vector space dimensionality and conform the feature set, we asked three sovereign financial experts to pre-filter the initially determined features. Afterwards we evaluated their selections and consider only noun phrases, which received concordant votes. So we establish a reference list of principal indicators representative for the fiscal policy domain.

3) Classifier Setup: The annotated documents, which contain the features and the corresponding instances, serve as a training set for the classification algorithm. Similar to the features, their values require also pre-filtering to omit the irrelevant. For this we execute a value selection task based on linguistic and statistical measures. The experiments are conducted with the simple term frequency, TFxIDF and the C-Value. Consequently, the candidate lists are first ranked and then trimmed using an adjustable threshold. The preliminary ten-fold cross validation implies that the C-Value weighting outperforms the other two methods by a small margin. Further tests with alternative thresholds and a neuronal network parser are in the process.

Since we consider all identified instances equally important for the learning model, we favor a multi-instance training approach. A multi-instance classifier is a supervised classification technique, which supports several sub-instances per attribute instance. Furthermore, all sub-instances are assigned only to one class. The corresponding format [14] is defined by three elements: a **bag-id** (the nominal attribute), a **bag** (the corresponding relational attribute with all instances), and a **class label**. We conduct the initial tests with the multi-instance version of the Sequential Minimal Optimization [15] algorithm. In order to enable a comprehensive evaluation we plan also experiments with standard single-instance classifiers.

The feature-set matrix consists of the extracted feature-value paris and a binary class. On each event day—where an official announcement is issued by the FED—we track also particular market movements. The initial tests are conducted with the S&P500 index, which comprises five hundred companies' stocks on the basis of factors such as liquidity and market size. Because of its composition the market value weighted index is considered as a benchmark of the entire U.S. stock exchange. Likewise it is a long term economy indicator, which incorporates future return expectations. Here, the class value is calculated by measuring the ratio—positive or negative—between the closing and the opening value on event days.

4) Model Evaluation: The study evaluation is fulfilled in two steps. First, we examine the information extraction approach. One metric applies to the feature extraction, where we empirically compare terminology extraction algorithms, i.e., C-value [13], GlossEx [16], and TermExtractor [17], with a reference subset of noun phrases, which is pre-filtered by domain experts. Here, C-Value scored best. The second metric involves the selection of the feature instances, which are labeled by a grammatical dependency parser, i.e., OpenNLP [18] and Stanford [19]. Hereby we compare the two parsers' outcome with manually annotated phrases. Our preliminary tests indicate, that models trained on the *WSJ* corpus excel. Whereas, Stanford's *RNN* [20] and *english-Factored* [19] statistical parsers perform similar.

In a second step, we evaluate our model by measuring the classifier performance using ten-fold cross-validation. Initially we experiment with the filter methods—TF, TFxIDF, C-Value—and the tresholds, which are independent from the learning model. In addition, the four different document categories are analyzed separately. Furthermore, we intend to compare our results with the state-of the-art metrics of similar studies.

The research work is currently in progress and the outlined architecture is a subject to minor updates. The quantitative and qualitative evaluation is still in progress. We conducted our first experiments with the presented model and the results are promising, although preliminary and incomplete.

IV. Conclusion

In this study we presented a novel approach to quantify financial news articles. We used linguistic and statistical measures to identify a set of domain relevant features and to extract the corresponding candidate values (instances). Consequently, we were able to determine principal indicators associated to the economy status in the US, which are incorporated in the central bank's announcements. As a result, our model enabled an explicit tracking of the monetary policy conducted by the Federal Reserve in conjunction with abnormal stock market movements. Furthermore, it facilitates economic surveys by providing a model for extracting financial variables and their conditional values.

Nevertheless, the model is limited to the financial domain. Empirically, all written texts are ambiguous and thus difficult to annotate automatically. This may be an indication why parsing algorithms struggle to keep a high cross-domain performance. In addition, the identification of decisive features for classification remains challenging, due to the rich structure of the natural language. Future works on news quantification should benefit from a supplementary field expertise. Capturing the semantics and the syntax of a particular domain is a likely start.

ACKNOWLEDGMENT

The author would like to thank for the contribution and the support to his research advisors: Prof. Christoph Schommer from the Computer Science and Communications Research Unit at the University of Luxembourg, Prof. Theoharry Grammatikos from the Luxembourg School of Finance, and Dr. Ulrich Schaefer from the Language Technology Lab at the German Research Center for Artificial Intelligence (DFKI) in Saarbuecken. And also to his family, friends and colleagues, who encouraged this work. This research paper would not be possible without all of them.

REFERENCES

[1] M. Minev, C. Schommer, and T. Grammatikos, "News and stock markets: A survey on abnormal returns and prediction models," University of Luxembourg, Tech. Rep. UL-Article-2013-018, Aug 2012.

[2] B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam, "Daily prediction of major stock indices from textual www data," in *KDD*, 1998, pp. 364–368.

[3] D. Peramunetilleke and R. K. Wong, "Currency exchange rate forecasting from news headlines," *Aust. Comput. Sci. Commun.*, vol. 24, no. 2, pp. 131–139, Jan. 2002. [Online]. Available: http://dx.doi.org/10.1145/563932.563921

[4] M. Hagenau, M. Liebmann, M. Hedwig, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-specific features," *Hawaii International Conference on System Sciences*, vol. 0, pp. 1040–1049, 2012.

[5] S. S. Groth and J. Muntermann, "An intraday market risk management approach based on textual analysis," *Decision Support Systems*, vol. 50, no. 4, pp. 680 – 691, 2011.

[6] B. Drury, L. Torgo, and J. Almeida, "Classifying news stories to estimate the direction of a stock market index," in *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on*, june 2011, pp. 1 –4.

[7] R. P. Schumaker and H. Chen, "A quantitative stock prediction system based on financial news." 2009, pp. 571–583.

[8] FED, "Federal reserve bank: Monetary policy," http://www.federalreserve.gov/monetarypolicy/default.htm, July 2013.

[9] A. Kurov, "Investor sentiment and the stock market's reaction to monetary policy," *Journal of Banking and Finance*, vol. 34, no. 1, pp. 139 – 149, 2010.

[10] C. Rosa, "How 'Unconventional' are Large-Scale Asset Purchases? The Impact of Monetary Policy on Asset Prices," *SSRN eLibrary*, 2012.

[11] J. Hausman and J. Wongswan, "Global asset prices and fomc announcements," *Journal of International Money and Finance*, vol. 30, no. 3, pp. 547 – 571, 2011.

[12] J. Sager, D. Dungworth, P. McDonald, and P. McDonald, *English special languages: principles and practice in science and technology*. Brandstetter, 1980.

[13] K. T. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the c-value/nc-value method," *Int. J. on Digital Libraries*, vol. 3, no. 2, pp. 115–130, 2000.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[15] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998. [Online]. Available: http://research.microsoft.com/jplatt/smo.html

[16] L. Kozakov, Y. Park, T.-H. Fin, Y. Drissi, Y. N. Doganata, and T. Cofino, "Glossary extraction and utilization in the information search and delivery system for ibm technical support," *IBM Systems Journal*, vol. 43, no. 3, pp. 546–563, 2004.

[17] F. Sclano and P. Velardi, "Termextractor: a web application to learn the shared terminology of emergent web communities," in *IESA*, 2007, pp. 287–290.

[18] "Apache openlp library," http://opennlp.apache.org/, August 2013.

[19] D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," in *Advances in neural information processing systems*, 2002, pp. 3–10.

[20] R. Socher, C. C. Lin, A. Ng, and C. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 129–136.