Economic Prediction using Heterogeneous Data Streams from the World Wide Web

Abby Levenberg[1], Edwin Simpson[2], Stephen Roberts[1,2], and Georg Gottlob[1,3]

[1] Oxford-Man Institute of Quantitative Finance,
University of Oxford,
abby.levenberg@oxford-man.ox.ac.uk
[2] Machine Learning Research Group,
Department of Engineering Science,
University of Oxford,
sjrob@robots.ox.ac.uk, edwin@robots.ox.ac.uk
[3] Department of Computer Science,
University of Oxford,
georg.gottlob@cs.ox.ac.uk

Abstract. Learning to predict financial and economic variables of interest is a hard problem with a large body of literature devoted to it. Of late there has been a significant amount of work on using sources of *text* from the Web (such as Twitter or Google Trends) to predict financial and economic variables. Much of this work has relied on some form or other of superficial sentiment analysis to represent the text. In this work we present a novel approach to predicting economic variables using multiple heterogenous streams of Web data. We can incorporate different data types into our model – such as time series and text – by first treating each data stream as a separate source with its own features and predictive distribution. For the text data streams we use a novel approach to prediction using a sentiment composition model to generate features. We then use a Bayesian classifier combination model to combine the independent "weak" predictions into a single prediction of the Nonfarm Payroll index, a primary economic indicator. Our results show that using a classifier combination model over multiple streams can achieve very high predictive accuracy.

Keywords: heterogeneous data streams, economic prediction, classifier combination, text sentiment

1 Introduction

There is a vast amount of data available on the Internet from a huge number of distinct online sources and the rate of its output is increasing daily. Currently there is significant interest in both industrial and academic research that aims to utilize such *big data* provided by the WWW to make predictions and gain insights into various aspects of daily life. Of late there has been a lot of work using textual WWW data to make predictions of a financial nature attempting to find correlations between the data and various lead economic and financial indicators such as the stock market or employment rates. Structured extraction of and learning from these online sources of data is a useful and challenging problem that spans the machine learning, information extraction, and quantitive finance research communities.
In this work we forecast the trend of the United States *Nonfarm Payrolls* (NFP), a monthly economic index that measures employment growth (decay) and is considered an important indicator of the welfare of the U.S. economy.[4] The NFP index is part of the Current Employment Statistics Survey, a comprehensive report released by the United States Department of Labor, Bureau of Labor Statistics, on the state of the national labor market. Released on the first Friday of each month, the index is given as the *change* in the number of (nonfarm) employment compared to the prior month. Besides indicating the state of the economy, the NFP is an index that "moves the market" upon its release [17] with the market reacting positively to a increase in the index and negatively to a decline. It is of interest to anyone with an stake in the market, such as banks, hedge funds, prop traders, etc., to try

[4] http://research.stlouisfed.org/fred2/series/PAYNSA?cid=32305

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

and make an accurate and timely prediction of its direction. As such, as the NFP release data nears there is a significant amount of speculation in the media from economists attempting to forecast its direction and value.

We show that such a prediction is possible using freely available data from the WWW. We present a novel extraction and machine learning framework to access and combine features from heterogeneous *data streams* from disparate online sources. We make use of both text and real-valued streams. For the text streams we present a novel approach to prediction using features generated from a state-of-the-art sentiment composition algorithm. We combine these text streams with relevant timeseries data mined from the WWW. We use these streams to learn accurate predictions of the future trend of our economic variable of interest. Since each stream provides its own predictive distribution we show how to fully exploit the information from separate streams of various data types by using an Independent Bayesian Classifier Combination (IBCC) model to obtain high accuracy in our predictive task. Using features from multiple WWW streams is a contribution to the current literature and presents a number of challenges that we address in the following sections.

In the next section we review the relevant literature in the area of using WWW data to make economic predictions. In Section 3 we present a stream-based framework for online extraction for multiple unrelated heterogeneous data sources. We also describe the IBCC model which enables us to aggregate any number of stream specific base classifiers into a single prediction. In Section 4 we describe the data streams in further detail. In Section 5 we report on correlating our data streams with economic trends and using the complete streaming framework with the IBCC model to predict the NFP. We show results that are state-of-the-art.

2 Prior Work

Much previous work has concentrated on the combination of information sources of the same type. In this paper, however, we combine heterogeneous data streams, time series and textdata, to achieve robust prediction models. Our review is therefore divided into two categories of prior work: time series and text data.

2.1 Time Series Prediction

Arguably the whole field of quantitative financial analysis revolves around the ability to detect signals within and between time series data. As such there is a large body of literature on using time series data to predict financial variables of interest. Techniques range from simple heuristics based on intuition and market knowledge to state-of-the-art machine learning algorithms such as genetic algorithms and deep networks. Over the last decades many textbooks have been and continue to be published that describe a huge number of techniques for finding correlations between various financial and economic time series, for example [3], [8], and [11]. Numerous journals and conferences are devoted to disseminating the latest approaches for financial timeseries prediction and regression for analysts, traders, quants and academics. For example, the *Journal of Time Series Econometrics* and the *Journal of Time Series Analysis* are journals devoted entirely to publishing the latest findings in this area.

2.2 Text Prediction

Utilizing the information implicit in market news and opinion to predict the direction of the economy is of obvious interest to many people. As such there has been significant amount of work that uses text from various online sources for prediction of economic indexes and stock market trends (see [9, 18, 5] for instance). In general the framework of these papers is to obtain natural language text from the Web, such as news stories, message board data, Twitter feeds, etc., and to use language specific features, often sentiment based, to train a classification algorithm to predict the future direction or value of the index/market. Learning algorithms range from simple two-class Naive Bayes and Support Vector Machines to more sophisticated algorithms with varying results and claims.

An overview and comparison of a number of such predictive systems tailored specifically to the stock market is given in [12] and [14]. Some of the reported work describes trading strategies based on system predictions that perform well beyond market expectations. However, the authors suggest the systems

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

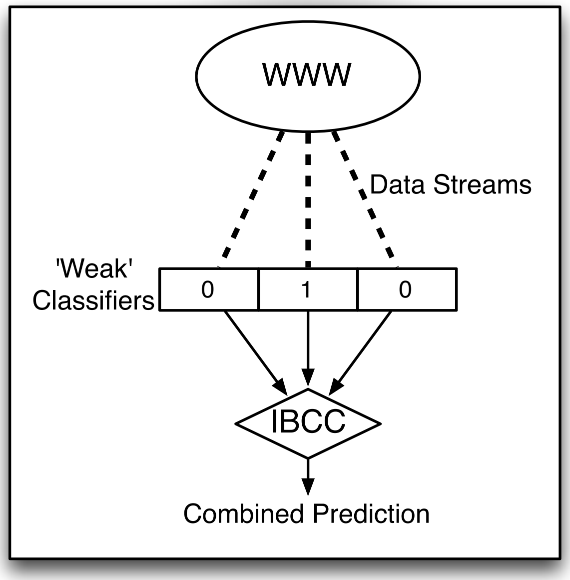

Fig. 1. Framework for prediction. We aggregate independent predictions from multiple heterogeneous data streams from the WWW into a single combined trend prediction using a Bayesian combination framework.

they review suffer from a lack of proper testing and unrealistic market expectations. As well, most of the systems reviewed in these summaries use a “bag of words” model to compute the features for the document-level classification. The authors argue this approach is far too general and accuracy from this is impacted due to the loss of contextual information from each document.

Current work has focussed on the use of big textual data to predict economic and market trends (see [4], [10], [16]). An example of note is [1]. Here the authors regressed from multidimensional sentiment moods (i.e., “Calm”, “Happy”) obtained from a stream of Tweets to the market and found some weak correlation with a single dimension of sentiment. While this research generated a significant buzz in the media and financial sectors its application to real-world trading remains unclear.

Other interesting work using text features for various predictions include the work described in the overview from [20] and [2]. Here a group of “text-driven forecasting” models are described that are used to predict phenomenon ranging from the volatility of yearly returns from financial reports, box office revenues from film critics reviews, and menu prices from the sentiment of costumer restaurant reviews. More recently [15] used *Google Trends* to find more significant correlations with changes in volumes of search queries of particular finanical terms and the lagged market trend.

3 Streaming Prediction Framework

Our goal is to efficiently use the big data freely available on the WWW to make predictions of economic variables of interest. However, for a given domain there is an overwhelming amount of data available from any number of sources. A simplifying conceptual approach for making sense of the abundance of Web data is to treat each online source of data as a separate *stream* of data. Each data stream has its own underlying distribution and throughput, the rate at which the source produces raw data, and hence its own independent level of predictive accuracy. If we treat each stream as a classifier in its own right we can make use of ensemble methods to combine the independent predictions into a single best prediction. As well, since each stream is considered independently this approach enables us to fuse multiple heterogeneous sources of data together into a single model of prediction. In this section we describe a framework for data stream extraction and aggregate prediction using IBCC from independent "weak" classifiers built from multiple WWW streams.

As Figure 1 depicts our framework for stream-based prediction is divided into three parts:

1. Extracting the relevant streams from the Web in a structured and efficient manner.
2. Training an ensemble of base classifiers – one for each data stream – using features and models specific to each stream.
3. Aggregate multiple, stream-specific classifications into a single globally optimised prediction.

Below we describe in further detail each part of this framework.

3.1 Structured Stream Extraction

Since we aim to predict the trend of the economic index, the NFP, we want to find streams that contain useful information for predicting the economy. An immediate question we must answer is how to find and extract *only* the data relevant to the predictive task at hand from the massive amount of data available online. Consider that even within a stream from a single source there may be data that pertains to an arbitrary number of domains. For example, a stream of text from a website that broadcast news in real-time will contain stories ranging from the economy to celebrity surgery and everything in-between. We may want to use the pertinent articles on the economy from such a source but indiscriminate collection of the stream will mean most of the text we collect will be irrelevant to our predictive task.

Hence we use a mechanism based on *Oxpath*, a query language for web data extraction that enables the automation of user-driven queries of a given source and then structured retrieval of the returned data [6]. For instance, suppose we aim to collect articles pertaining to the NFP from the online archives of various newspapers and magazines. Using Oxpath we can setup an automated process to periodically query multiple sources for particular terms over specific dates, daily for instance, and save the data returned as structured entries into a local repository. This enables us to capture details present on a web page such as the author, title, date, etc., of an article. This means we do not have to download, process, and classify raw HTML pages from the web which is a tedious and error prone process. Instead we have direct structured access to the desired content of the stream.

3.2 Stream-specific Classifiers

Once we have access to the pertinent data from a particular data source we need to train a predictive model specific to that stream to forecast the NFP, our dependent variable. Here any of the standard machine learning models in the literature are viable. For example, since we are predicting the directional trend of an economic index we use simple binary logistic regression models where a class of 1 means "up" and 0 means "down". However, to use any predictive models we first must derive features from the raw streams to use as training data to our classifier.

In this work we use both real-valued time series and text data streams. For the time series data we can use standard multivariate features such as smoothed moving averages etc. For the text data we try something simple but new. First we use sentiment composition to score individual sentences with a distribution over positive, negative, or neutral sentiment [13]. Afterwards we combine these sentence-level sentiment features in some informative way as input into our training algorithms. In Section 5 we report experiments on various approaches for combining the sentiment distribution from individual sentences as input features for model training. Next we describe how we combine these stream-specific predictions into a single best prediction.

3.3 Binary IBCC Model

Due to the differences in their underlying distributions, each of the individual data stream's predictive accuracies may very enormously in reliability. Classifier combination methods are well suited to situations such as these and serve to make best use of the outputs of an ensemble of imperfect base classifiers to enable higher accuracy classifications. Using a *Bayesian* approach to classifier combination provides a principled mathematical framework for aggregation where poor predictors can be mitigated and in which multiple data streams, with very different distributions and training features, can be combined to provide complementary information [7]. Here we describe a binary, two-class variation of the IBCC model of [19]. [5]

[5] The full model for an arbitrary number of classes ≥ 2 is described in [19].

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

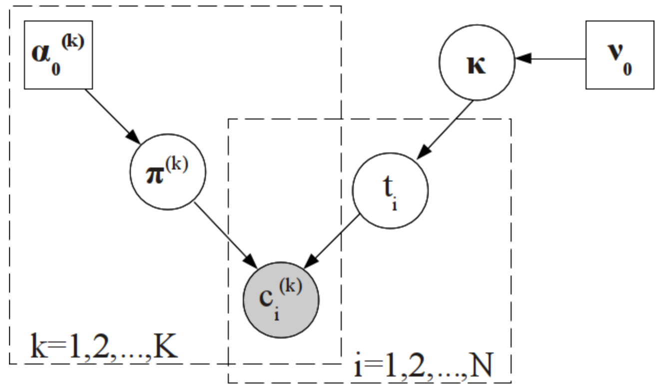

Fig. 2. Graphical model for IBCC. The arrows indicate dependencies while the shaded node represents observed variables, the square nodes are hyper-parameters. All other variables must be inferred. Here the predictions $\boldsymbol{c}_i^k$ of each base classifier k are generated dependent on the confusion matrices $\boldsymbol{\pi}^k$ and the true label t_i.

We want to predict the trend of the NFP over some number of months, or more generally *epochs*, indexed from $i \in \{1, \ldots, N\}$. We assume the trend $\boldsymbol{T}$ of the NFP is generated from an underlying binomial distribution with parameters $\boldsymbol{\kappa}$. Each epoch has a value $t_i \in \{0, 1\}$ where the ith epoch has a label $t_i = 0$ if the NFP index decreased from the prior epoch and $t_i = 1$ if it increased. The prior probabilities of the trends t_i are given by $\boldsymbol{\kappa} : p(t_i = j|\boldsymbol{\kappa}) = \kappa_j$, where j iterates over the class labels $\{0, 1\}$.

We denote the number of base classifiers, or data streams, as K. Each stream's base classifier $k \in \{1, \ldots, K\}$ produces a real-valued output matrix $\hat{\boldsymbol{C}}^k$ of size $N \times j$. The output vector $\hat{\boldsymbol{c}}_i^k \in [0, 1]$ for epoch i denotes the probabilities given by classifier k of assigning a discrete trend label $\boldsymbol{c}_i^k \in \{0, 1\}$. The jth element of the trend label, $c_{ij}^k = 1$, while all other elements are zero, indicates that classifier k has assigned label j to epoch i. We assume the vector $\boldsymbol{c}_i^k$ is drawn from a binomial distribution dependent on the true label t_i, with probabilities $\boldsymbol{\pi}_j^k = p(\boldsymbol{c}_i^k|t_i = j, \boldsymbol{\pi}_j^k)$. Both parameters $\boldsymbol{\pi}^k$ and $\boldsymbol{\kappa}$ have Beta-distributed priors.

The joint distribution over all variables for the binary IBCC model is

$$p(\boldsymbol{\kappa}, \boldsymbol{\Pi}, T, \boldsymbol{C}|\mathbf{A}_0, \boldsymbol{\nu}) = \prod_{i=1}^{N} \{\kappa_{t_i} \prod_{k=1}^{K} \boldsymbol{\pi}_{t_i}^k \cdot \boldsymbol{c}_i^k\} p(\boldsymbol{\kappa}|\boldsymbol{\nu}) p(\boldsymbol{\Pi}|\mathbf{A}) \tag{1}$$

where $\boldsymbol{\Pi} = \{\boldsymbol{\pi}_j^k | j \in \{1, 0\}, k = 1 \ldots K\}$ denotes all base classifier probabilities, $\mathbf{A}_0 = \{\boldsymbol{\alpha_{0j}}^k | j \in \{1, 0\}, k = 1 \ldots K\}$ the corresponding set of hyper-parameters, and $\boldsymbol{\nu_0} = [\nu_0, \nu_1]$ are the hyper-parameters for $\boldsymbol{\kappa}$. A graphical model of IBCC is shown in Figure 2.

The probability of a test point t_i at epoch i being assigned class j is given by

$$p(t_i = j) = \frac{\rho_{ij}}{\sum_{y=1}^{J} \rho_{iy}} \tag{2}$$

where

$$\rho_{ij} = \kappa_j * \prod_{k=1}^{K} (\boldsymbol{\pi}_j^k \cdot \boldsymbol{c}_i^k) \tag{3}$$

which accounts for the probability of the class κ_j weighted by the combined prediction probabilities $\boldsymbol{\pi}_j^k$ of each stream's independent predictions $\boldsymbol{c}_i^k$.

A key feature of IBCC is that each base classifier k is modelled by $\boldsymbol{\pi}^k$, which intuitively represents a *confusion matrix* that quantifies the decision-making abilities of the indivdual base classifier k. The goal of inference for the model is to optimise the distributions over the unknown variables $\boldsymbol{T}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\kappa}$ such that the probability of t_i for each epoch i is maximized for epochs with true increases in the NFP and minimised for epochs i where the NFP decreased. In [19] this approach has been shown to outperform a number of baseline combination methods for classification tasks.

Positive Sentiment Example

"The Governor noted that despite jobs being down, there was a surprising bright spot: construction added 1,900 jobs in November - its largest gain in 22 months."

positive: *0.925* negative: *0.0* neutral: *0.075* confidence: *0.69*

Negative Sentiment Example

"When I drive down the main street of my little Kansas City suburb I see several dark empty storefronts that didn't used to be that way."

positive: *0.0* negative: *0.973* neutral: *0.027* confidence: *0.67*

Mixed Sentiment Example

"We continue to far better than the nation - our rate has been at or below the national rate for 82 out of the past 83 months - but we must also recognize that there were 10,2000 job lost at the same time."

positive: *0.372* negative: *0.591* neutral: *0.037* confidence: *0.73*

Fig. 3. Examples of sentences with different sentiment distributions accounting for the positive, negative, and neutral dimensions of a sentence.

4 Data

In this section we describe the data we collected to test our streaming prediction framework. Our timeline for training and testing spanned the NFP index monthly from January, 2000 through December, 2012. We made use of both time series data and text data from a variety of online sources described below.

We collected time series data online from a variety of online sources including the Federal Reserve Economic Data website [6], a Federal Reserve bank resource which compiles and maintains a large number of economic time series and data sets. Other sources included the Bureau of Labor Statistics [7] and the Conference Board [8] which both publish various economic indexes. We collected 33 different time series from such online sources to use as independent variables for predicting the NFP. We describe in detail our predictive models in the following section.

We also collected a number of textual data streams from multiple sources. For this we ran pointed queries against a large news database [9] and collected archived test data from nearly 700 distinct online text sources such as the Associated Press, Dow Jones, Wall Street Journal, etc. Altogether we collected over 6.6 million sentences of raw text from the streams.

After we collected the text data we processed the text at the sentence level for individual sentiment analysis using the model in [13] [10]. After sentiment analysis each sentence is represented as a distribution over three dimensions of sentiment: positive, negative, and neutral. Figure 3 shows some example results from the sentiment analysis system. In the next section we detail our experiments for prediction based on text, timeseries and their combination. using these sentiment dimensions as features.

5 Experiments

5.1 Experiment Setup

As described in Section 4 we collected data over a timeline of 13 years from 2000-2013 which contained 156 monthly epochs. We used the last 24 epochs as test points and the rest of the epochs in the timeline

[6] `http://research.stlouisfed.org/fred2/`
[7] `http://www.bls.gov/`
[8] `http://www.conference-board.org/`
[9] `http://www.dowjones.com/factiva/index.asp`
[10] The sentiment model we used is available as an API service from `http://theysay.io/`.

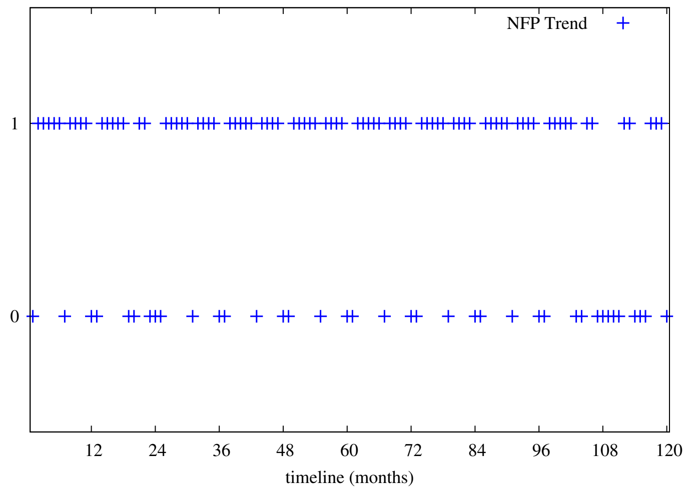

Fig. 4. The full NFP index trends over a 10 year timeline where 1=up, 0=down.

Source	AUC
Random	0.52
Always Up	0.50
Back Returns	0.54

Table 1. Baseline results for predicting the NFP index.

as training points. However, as the economy normally tends towards growth, save for in periods of recession, there is an over representation of 109(70%) positive cases compared to only 47(30%) negative instances in the NFP index since 2000. This is shown clearly in Figure 4 where the 1 on the y-axis indicates an upward trend of the NFP index and a 0 indicates a downward trend. To ascertain whether our approach is valid for learning good predictions rather than just optimising for the overrepresented class we subsampled randomly from the positive class to obtain a balanced training set.
For each data source we learnt a base classifier independently and used *rolling* predictions so the features associated with a given test point became part of the training data for the next test epoch. These models were then used as base inputs for IBCC. Note that the stream-specific classifiers need not give good individual prediction results as long as each contain useful information for the IBCC model. In fact base classifiers with very poor accuracy may be useful as IBCC can account for negative results so long as there is consistent information encoded in the probabilities. We measure our results using the standard metric Area Under the Receiver Operating Characteristic Curve (AUC). The AUC is the probability of ranking a positive example higher than a negative example and takes into account both true and false positive predictions [21]. For completeness, in the results that follow we show the results for the individual streams as well as the results using the IBCC model.

5.2 Baselines

Table 1 reports some baseline measures of prediction standard for the NFP. *Random* uses a random number generator to output a real number between $[0, 1]$ which it treats as prediction probabilities. *Always Up* always predicts the NFP as rising with a probability of 1. We also used the industry standard of *Back Returns* and predict each epoch will follow the trend of the last. Each of these achieve an AUC around 0.5 - as expected since subsampling makes the empirical priors of up and down equal and these methods do not have predictive power, with back returns performing at near random performance.

Source	Averages	Trends
Associated Press*	0.69	0.37
Dow Jones	0.44	0.25
Reuters News	0.46	0.36
Market News Intl.*	0.70	0.23
Other Sources*	0.63	0.63
Wall Street Journal	0.63	0.53
IBCC	0.81	**0.85**

Table 2. Stream-specific and combined AUC results for predicting the NFP index. We get better prediction accuracy using multiple sources (starred) with IBCC.

Source	AUC
CPI	0.70
ISM	0.85
JOLTS	0.66
LFL	0.71
IBCC	**0.90**

Table 3. Stream-specific and combined AUC results for predicting the NFP index using time series data. Here again accuracy is improved when using IBCC.

5.3 Text Stream Prediction

In this section we report on results predicting the NFP using the text streams both as independent classifiers and as base inputs to the IBCC model. Our general approach to using the sentiment features described in Section 4 is to aggregate the sentiment distributions over all sentences in an epoch and then use this representation as feature input into a simple logistic regression classifier models.

For example, the first results column in Table 2 shows the results when we use the percentages of word-weighted positive versus negative sentiment for each epoch for NFP trend prediction. The third column of Table 2 presents results using all the dimensions of sentiment available but using the *differences* in the counts between epochs as features. The idea behind this approach is intuitive and assumes the trends of sentiment implicit in the text should correlate with the trends of the economy. A raised level of negativity in the news media compared to normal would reflect a period of economic difficulty and visa versa for positive sentiment in the news. We can see this approach achieves a good measure of correlation between the text sentiment and the trends of the NFP.

These results over text show clearly there is predictive information within economic news that we can access via selecting intuitive features from the sentiment analysis of the text. Using these sentiment features in a state-of-the-art machine learning framework gives good prediction results for the NFP that significantly beat the baselines.

5.4 Time Series Prediction

Using the same methodology as above we build a suite of independent classifiers based on the time series data we collected and described in Section 4. We collected over thirty different economic indexes but here we report only on the four series with the best independent prediction results: the Consumer Price Index (CPI), the Institute for Suppy Management Manufacturing Index (ISM), the JOLTS Nonfarm Index (JOLTS), and the Labor Force Levels (LFL). Each of these is directly or indirectly related to the unemployment rate and hence the NFP. As with the text streams, for each time series we trained a logistic regression classifer using multivariate features from the data. The features consisted of the point value plus a number of indicators of the trend and moving averages for various time frames.

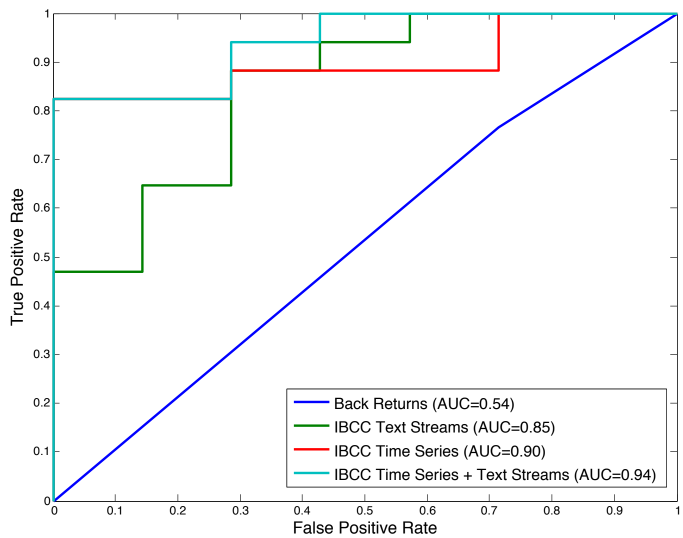

Fig. 5. The AUC results for the NFP predictions. Using a combination of text and time series data results in the best prediction accuracy.

As can be seen from Table 3 each individual time series gives significantly better results than the baselines and improves upon the text sentiment results. When we combine each of these weak classifiers using the IBCC model we get an improved overall AUC of 0.90.

5.5 Heterogeneous Data Prediction

Finally we tested combining the different data types – time series and text stream data – into a single prediction. This is straight forward since each data source is treated as an independent base classifier so IBCC cannot distinguish between the data types. As Table 4 shows, using both the data types together provides significantly improved prediction accuracy indicating that the sentiment within the text streams contains information that is complementary to the real-valued time series.

Source	AUC
Time Series + Text Averages	**0.94**
Time Series + Text Trends	0.91

Table 4. The AUC results when we combine hetereogenous data types with the IBCC model.

Figure 5 depicts the AUC results between the baselines and the IBCC results. Clearly we are learning something of interest using our streaming framework and associated combination model. As well we see the text data is providing us with a source of knowledge that is not present in the time series and, when used in a classifier combination setting, provides extra useful information that improves prediction. [11]

[11] To the authors' knowledge there is no prior published benchmark for NFP prediction against which to make a direct comparison. Our primary comparison here is against the stream-specific weak classifiers.

The Proceedings of ECML/PKDD 2013 Workshop
Scalable Decision Making: Uncertainty, Imperfection, Deliberation (SCALE)
September 23, 2013, Prague, Czech Republic

6 Conclusion

Using news streams and other text sources to make economic predictions is an area that has generated significant interest in the last decade. Our results show clearly there is predictive information within economic news that we can access via selecting intuitive features from the sentiment analysis of the text. Using these sentiment features in a state-of-the-art machine learning framework gives good prediction results of economic trends and variables of interest such as the NFP. However, our results show that combining these text streams with more standard time series data within a classifier combination framework such as IBCC produces highly accurate predictions. Clearly there is information within the text that is complementary to the information contained in the time series data. Using IBCC allows easy integration of multiple classifiers of arbitrary data types from a variety of sources and allows us to model the complementary information to obtain better results.
The scope of this type of economic prediction has many potential applications in both further academic research to more direct financial and market orientated ones with a host of directions for future work. For example, extending the classifier combination model to produce real-valued predictions instead of just predicting trend categories is research which we are current conducting. And while there is large scope for future work on using sentiment of big WWW text data for economic predictions, we believe the research we have reported in this paper is a step forward in the current literature in this area.

References

1. Bollen, J., Mao, H., Zeng, X.J.: Twitter mood predicts the stock market. J. Comput. Science 2(1), 1–8 (2011)
2. Chahuneau, V., Gimpel, K., Routledge, B.R., Scherlis, L., Smith, N.A.: Word salad: Relating food prices and descriptions. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1357–1367. Association for Computational Linguistics, Jeju Island, Korea (July 2012), `http://www.aclweb.org/anthology/D12-1124`
3. Chatfield, C.: The Analysis of Time Series: An Introduction. CRC Press (1975)
4. Choi, H., Varian, H.R.: Predicting the present with google trends. The Economic Record 88(s1), 2–9 (2012), `http://EconPapers.repec.org/RePEc:bla:ecorec:v:88:y:2012:i:s1:p:2-9`
5. Fan, D.P.: Predicting the index of consumer sentiment when it isnt measured. In: JSM Proceedings, AAPOR. p. 60986110. American Statistical Association, Alexandria, VA (2010)
6. Furche, T., Gottlob, G., Grasso, G., Schallhart, C., Sellers, A.: Oxpath: A language for scalable data extraction, automation, and crawling on the deep web. The VLDB Journal pp. 1–26 (2012), `http://dx.doi.org/10.1007/s00778-012-0286-6`
7. Ghahramani, Z., Kim, H.C.: Bayesian classifier combination. Gatsby Computational Neuroscience Unit Technical Report GCNU-T., London, UK: (2003)
8. Hamilton, J.D.: Time Series Analysis. Princeton University Press (1994)
9. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Mining of concurrent text and time series. In: In proceedings of the 6 th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining Workshop on Text Mining. pp. 37–44 (2001)
10. Mao, H., Counts, S., Bollen, J.: Predicting financial markets: Comparing survey, news, twitter and search engine data. CoRR abs/1112.1051 (2011)
11. Mills, T.C., Markellos, R.N.: The Econometric Modelling of Financial Time Series. Cambridge University Press (2008)
12. Mittermayer, M.A., Knolmayer, G.: Text mining systems for market response to news: A survey. Proceedings of the IADIS European Conference Data Mining (2007)
13. Moilanen, K., Pulman, S.: Sentiment composition. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2007). pp. 378–382 (September 27-29 2007), `http://users.ox.ac.uk/~wolf2244/sentCompRANLP07Final.pdf`
14. Nikfarjam, A., Emadzadeh, E., Muthaiyah, S.: Text mining approaches for stock market prediction. In: Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. vol. 4, pp. 256–260 (2010)
15. Preis, T., Moat, H.S., Stanley, H.E.: Quantifying trading behavior in financial markets using google trends. Scientific Reports 3(1684) (April 2013)

16. Preis, T., Reith, D., Stanley, H.E.: Complex dynamics of our economic life on different scales : insights from search engine query data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences Vol.368(No.1933), 5707–5719 (2010), http://wrap.warwick.ac.uk/50936/
17. Savor, P., Wilson, M.: How much do investors care about macroeconomic risk? evidence from scheduled economic announcements. Journal of Financial and Quantitative Analysis FirstView, 1–62 (3 2013)
18. Schumaker, R.P., Chen, H.: Textual analysis of stock market prediction using breaking financial news: The azfin text system. ACM Trans. Inf. Syst. 27(2), 12:1–12:19 (2009)
19. Simpson, E., Roberts, S., Psorakis, I., A., S.: Dynamic bayesian combination of multiple imperfect classifiers. Decision Making and Imperfection. Intelligent Systems Reference Library 474 (2013)
20. Smith, N.A.: Text-driven forecasting. http://www.cs.cmu.edu/~nasmith/papers/smith.whitepaper10.pdf (2010), http://www.cs.cmu.edu/~nasmith/papers/smith.whitepaper10.pdf
21. Spackman, K.A.: Signal detection theory: valuable tools for evaluating inductive learning. In: Proceedings of the sixth international workshop on Machine learning. pp. 160–163. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1989)