Forecasting Consumer Spending from Purchase Intentions Expressed on Social Media

Viktor Pekar and **Jane Binner**
Business School, University of Birmingham, Birmingham, UK
{v.pekar, j.m.binner}@bham.ac.uk

Abstract

Consumer spending is a vital macroeconomic indicator. In this paper we present a novel method for predicting future consumer spending from social media data. In contrast to previous work that largely relied on sentiment analysis, the proposed method models consumer spending from purchase intentions found on social media. Our experiments with time series analysis models and machine-learning regression models reveal utility of this data for making short-term forecasts of consumer spending: for three- and seven-day horizons, prediction variables derived from social media help to improve forecast accuracy by 11% to 18% for all the three models, in comparison to models that used only autoregressive predictors.

1 Introduction

Social media is increasingly reflecting many social phenomena that previously could be studied only with traditional surveying techniques such as telephone or face-to-face interviews. Recent research has demonstrated that it can be used to track the spread of epidemics (Culotta, 2010), monitor mass emergency situations (Nguyen et al., 2017), study political preferences during election campaigns (Tumasjan et al., 2010), predict product sales (Elshendy et al., 2017) and stock price changes (Si et al., 2014).

In this paper we examine the idea that social media can provide useful evidence about consumer confidence, a macroeconomic indicator describing the propensity of households to consume goods and services in the near future. Consumer confidence is one of the most crucial indicators of the health of an economy, as consumer spending constitutes the largest component of GDP in many developed countries. Government institutions and market research agencies compile their consumer confidence indices on a regular basis. Among the best-known ones are the Consumer Sentiment Index produced by University of Michigan for the US and GfK's Income Expectation and Willingness-to-buy indicators for the EU. These measures are obtained using traditional surveys, which have significant drawbacks: they are costly to conduct, based on low-frequency observations and published with substantial delays. Social media data hold the promise to overcome these drawbacks.

Previous research studied models of consumer spending trained on search engine data, based on the intuition that web searches for product names indicate intended purchases (Vosen and Schmidt, 2011; Scott and Varian, 2015; Wu and Brynjolfsson, 2015). Search engine data, however, do not capture the context of the purchase intention, such as the context available on social media in the form of extended coherent text, and thus are more likely to contain noise. A number of studies aimed to estimate a consumer confidence index from social media using sentiment analysis (O'Connor et al., 2010; Daas and Puts, 2014; Igboayaka, 2015). These methods derive a sentiment index from messages related to the economic outlook, which is compared with an official index to detect correlation or to train a model to predict it.

In contrast to this work, our method aims to model future consumer spending from purchase intentions expressed on social media. The method determines phrases referring to intended purchases and creates their condensed semantic representations, which are then used in a regression model alongside autoregressive predictors. Our experiments with time series analysis models (Seasonal Autoregressive Integrated Moving Average) and

Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–101
Copenhagen, Denmark, September 7–11, 2017.

machine-learning regression models (AdaBoost and Gradient Boosting) demonstrate utility of this data for making short-term forecasts of consumer spending. We find that for three- and seven-day horizons the semantic predictors help to improve forecast accuracy by 11% to 18% for all the three models.

The main novel contributions of this paper are (i) a prediction model that uses semantic information obtained from purchase intentions, which allows on the one hand, to abstract from specific lexical data, and on the other, reduce the complexity of the model; (ii) a study of optimal forecast horizons for the model that uses this information; (iii) an investigation of possibilities to incorporate semantic predictors with endogenous variables (i.e., lagged values of the consumer spending index) within the model.

The remainder of the paper is organized as follows. In the next section we review related work. The proposed method is described in Section 3. Section 4 details experimental setup. Results and their discussion are presented in Section 5. Section 6 concludes.

2 Related work

2.1 Sentiment analysis

A popular approach in previous work on modelling economic indicators from textual data has been to use automatically detected sentiment of documents. The study by O'Connor et al. (2010) predicts consumer confidence from sentiment found in Twitter posts that contain pre-defined keywords, such as "economy" or "job". Sentiment is assessed using a lexicon-based method and a daily sentiment index is constructed, which is then used as a predictor in an ordinary least-squares model of the ICS index. Daas and Putz (2014) take a similar approach, using a commercial sentiment analyser and a list of economy-related keywords, to study consumer confidence in Dutch social media. They find their sentiment measure to correlate and co-integrate with an official consumer index. Georgoula et al. (2015) use time-series analysis to study the relationship between Bitcoin prices, fundamental economic variables, and measurements of collective mood derived from Twitter. Using an SVM classifier trained on tweets mentioning Bitcoin, they obtain a sentiment measure which is used as a variable in an OLS and a VECM models. Souza et al. (2016) examine the relationship between Twitter sentiment, on the one hand, and the trade volume, returns, and volatility of selected stocks, on the other. Their method uses a domain-independent SVM classifier to construct a daily sentiment index, which is then used in a VAR framework along with the economic variables. Granger causality tests are used to identify causality links between these variables.

2.2 Lexical analysis

Sentiment analysis is known to be a difficult NLP problem, where accuracy varies greatly depending on domain customization. Therefore, methods that use lexical information instead seem to be an interesting alternative. Dergiades et al. (2015) examined raw counts of Twitter and Facebook posts containing "Grexit"-related words, detecting causality from them to changes in Greek government bonds for the same time period using Granger causality tests. Scott and Varian (2015) use search engine queries as predictors of Consumer Sentiment Index. To deal with the "fat regression" problem (the number of potential predictors is similar or even greater than the number of available observations), they introduce a Bayesian method to select predictor variables.

To deal with a large number of predictors derived from lexical data, various dimensionality reduction techniques have been proposed. Coussement and Van den Poel (2008) predict customer churn from the text of call centre emails. Creating classification features using Latent Semantic Indexing applied to the email corpus, they combined them with features traditionally used to predict customer churn (such as product usage data) in a maximum entropy classifier, and found that the former were helpful in identifying customers prone to churn. Rönnqvist and Sarlin (2015) analyse news articles to predict "bank distress" events, such as government interventions. Their approach constructs para2vec (Le and Mikolov, 2014) representations of news articles which are input into a neural network model to predict a distress score for a bank.

2.3 Combining sentiment and lexical data

Several papers used a combination of sentiment and lexical information in their models. Hansen and McMahon (2016) assess the effect of central bank communications on different market and real economic variables. From a corpus of central bank publications, they estimate an LDA model

and manually select those topics that have to do with a discussion of economic outlook. A dictionary-based sentiment analysis is used to obtain a monthly sentiment index, which is input as a variable in a Factor-Augmented VAR framework. Archak et al. (2011) present a hedonic regression model of product sales that uses customer reviews of the products as input. The reviews are analysed to extract nouns as potential references to product features and adjectives related to the nouns as potential evaluative phrases. The noun-adjective co-occurrences are arranged into a matrix which is then transformed using a technique similar to ANOVA decomposition. The reduced dimensionality matrix is input as variables of a regression model, along with non-textual variables such as the price of the product. Si et al. (2014) use a combination of lexicon-based sentiment analysis and LDA topics extracted from Twitter posts containing a stock's ticker symbol, on which the stock's price is regressed using a VAR model.

3 Proposed method

Our method aims to predict an official consumer spending index from the mentions of purchase intentions. Specifically, we expect that the semantics of noun phrases that are stated as intended purchases will be predictive of the official index for a certain number of subsequent days. The method consists of the following steps. First, tweets mentioning a purchase intention are collected from Twitter API. Second, noun phrases referring to the objects of the intended purchases are extracted and their daily counts are obtained to create a noun-by-date matrix. In order to account for semantic similarities between the nouns, a word2vec model is used to create a semantic vector for each date. Finally, a regression model of the consumer index is trained that uses the semantic vectors as well as lagged values of the index. These steps are detailed in the following sections.

3.1 Detecting purchase intention

Prior work on recognizing intentions have used both rule-based (Hamroun et al., 2016) and machine learning approaches (Chen et al., 2013). In this paper we opt for a rule-based method, as it can ensure high precision, while recall is of a less concern considering large volumes of available data. To obtain tweets mentioning purchase intentions, we issue a set of queries to the Twitter Search API, which are meant to capture common ways to express an intention to buy something. They are created from combinations of (1) first-person pronouns (”I” and ”we”), (2) verbs denoting intentions (”will”, ”’ll”, ”be going to”, ”be looking to”, ”want to”, ”wanna”, ”gonna”), and (3) verbs denoting purchase (”buy”, ”shop for”, ”get oneself”), thus obtaining queries such as ”I will buy” or ”we are going to buy”.

The text of each tweet is cleaned (any material outside of the grammatical text is removed) and processed with a part-of-speech tagger. PoS tag patterns are then applied to extract the head noun of the noun phrase following the purchase verb (e.g., ”headphones” in ”I am looking to buy new headphones”). After that, daily counts of the head nouns are calculated.

3.2 Semantic vectors

To represent the semantics of the nouns, we use the word2vec method (Mikolov et al., 2013) which has proven to produce accurate approximations of word meaning in different NLP tasks (Baroni et al., 2014). A word2vec model is a neural network that is trained to reconstruct the linguistic context of words. The model is built by taking a sequence of words as input and learning to predict the next word, using a feed-forward topology where a projection layer in the middle is taken to constitute a semantic vector for the word, after connection weights have been learned. The semantic vector is a fixed-length, real-valued pattern of activations reaching the projection layer. For each word, the input text originally has a dimensionality equal to the vocabulary size of the training corpus (typically millions of words), but the semantic modelling provides reduction to the size of the vector (typically several hundreds). The reduced dimensionality helps to reduce the complexity of the models, prevent overfitting, and is beneficial in computationally intensive classification and regression algorithms.

For each date, we map each noun that was observed on that day to a semantic vector, using word2vec vectors trained on a large corpus of Twitter posts. The semantic vectors of all the nouns for each day are then averaged to obtain a single vector. The components of the vectors will then be used as variables in regression models.

To allow for some time between the stated purchase intention and the actual purchase, we exper-

iment with different numbers of days between the day on which intentions were registered and the day for which the value of the consumer spending index is predicted.

3.3 Combining endogenous and exogenous variables

Our method makes predictions based on endogenous variables (i.e., lagged values of the index itself) and exogenous variables (i.e., semantic vectors obtained from Twitter). Thus, given a target value of the consumer spending index y_t at day t, a lag p, a k-dimensional semantic vector, and allowing for s days between the day when purchase intentions were registered and the day for which spending was reported (i.e., day t), a training instance is composed of endogenous variables $y_{t-1}, y_{t-2}, ..., y_{t-p}$ and exogenous variables $x^1_{t-s}, x^2_{t-s}, ..., x^k_{t-s}$.

We implemented two ways to combine the two types of variables to obtain a prediction. The first is simple concatenation of the variables into one vector of predictors. The second involves first training separate regression models for the endogenous variables and semantic variables separately, and then using the predicted values of each to train a third model that outputs the final predicted value.

3.4 Regression methods

In our experiments we include the following regression methods[1].

SARIMA(X). The Seasonal Autoregressive Integrated Moving Average (SARIMA) is a variety of the general ARIMA model. ARIMA(p,d,q) is defined via terms p, d, and q, where p represents the number of time-lagged variables; d – the number of differences required to remove seasonality and make the forecast variable stationary; and q – the number of time-lagged error parameters to account for an observed moving average. The orders of p and q can be identified using an autocorrelation and a partial autocorrelation function, or using information criteria, such as Akaike IC, or estimated from a validation set. The degree of differencing can be determined using stationarity tests such as the Dickey-Fuller test. Given order values, coefficients of the model can be estimated by least square regression or maximum likelihood estimators.

[1]We use the implementations in the `scikit-learn` and `statsmodels` packages.

SARIMA is formed by including additional seasonal terms: SARIMA(p, d, q)(P, D, Q)$_m$, where P, D, and Q are used to represent seasonal autoregressive model, the degree of seasonal differencing, and the seasonal moving average, correspondingly, while m stands for the length of the seasonal period. To identify the P, D, Q, and m terms, the autocorrelation and partial autocorrelation algorithms or information criteria can also be used.

SARIMAX is a SARIMA that allows for one or more exogenous variables to be included into the regression. We input the semantic vector as exogenous variables into SARIMAX.

AdaBoost Regression. AdaBoost (Freund and Schapire, 1996) is a machine-learning ensemble algorithm that uses the entire training data to successively train a series of weak learners, such as decision stumps. After one weak model is trained, the algorithm identifies the most difficult instances and computes their weights to exaggerate their effect on the training of the next model. The objective of this step is to "teach" the next model to correctly predict the test instances on which errors were made. Initially all instances have the same weight and hence have the same impact on training of the initial model. After each iteration, the weights of instances are adjusted, while the weights of instances with accurate predictions are decreased. Furthermore, each model is assigned a weight based on its overall accuracy. During the testing phase, the forecast values and the weights of the models are taken into account to produce a weighted average value.

Gradient Boosting Regression. Gradient Boosting (Friedman, 2001) is a gradient descent ensemble algorithm, which, similar to other boosting methods, operates by sequential training of weak models, which collectively would form a strong model. This is accomplished by training successive regression models on the residuals of the previous model, computed from errors it made. With each training round, Gradient Boosting improves the previous model by adding to it a new model that is trained only on the residuals, thus gradually fixing up errors made in the previous steps. To prevent overfitting, we additionally use an early stopping technique: the training of the model stops, if the validation loss has been increasing in four consecutive iterations.

During evaluation, we experimentally deter-

mine parameters of AdaBoost and Gradient Boosting on a validation dataset using the grid search technique. The model with the best parameter configuration is then evaluated on the test set.

4 Experiment setup

4.1 Data

Consumer Spending Index. As the forecast variable in our model, we use the Gallup Consumer Spending Index (CSI) [2]. The index represents the average dollar amount Americans report spending on a daily basis. The survey is conducted using telephone interviews with approximately 1,500 national adults. Respondents are asked to reflect on the day prior to being surveyed and provide an estimate of how much money they spent on that day. The eventual index is presented as a 3-day and a 14-day rolling averages of these amounts. In our evaluation, we used the 3-day values of CSI, between October 1, 2015 and July 31, 2016, i.e. 297 days in total.

Twitter. For the same period, we collected Twitter posts that originate from the US and that express intentions to buy, obtaining the total of 68,730 messages. Counts of nouns referring to purchases were extracted and rolling averages for each noun for three-day periods were calculated. To eliminate noisy data, we selected the 1000 most common nouns to construct semantic vectors.

Semantic vectors. Considering the amount of available training instances, we use the 25-dimensional vectors pre-trained on a large corpus of Twitter posts from the GloVe project[3].

Train-validation-test split. The available data was divided into the training, validation and test parts, in proportion 60%-20%-20%. The CSI values and their split into the three parts are shown in Figure 1. Because we use seven-day lags to create endogenous variables, there are seven-day gaps between the train and validation sets as well as between the validation and test sets there are seven day gaps, to ensure that no training data is used for validation or testing.

4.2 Evaluation method

Once a model was trained on the training set and its parameters optimized on the validation set, it was evaluated on the test set using dynamic forecasting: given the first day t of the test set, and the forecast horizon h, the model predicted h days in the future, for each day from t_2 to t_h the values predicted by the model for previous days were input as endogenous variables. In the following, we report results for h = 1, 3 and 7.

As evaluation metric, we use the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{T}\sum_{n=1}^{T}(y_n - \hat{y}_n)^2}$$

where y_n and $\hat{y}_n$ are the actual observation and the predicted value at day t_n, and T is the set of test values.

As the baselines, we use prediction models trained with the same algorithms but only on endogenous variables.

5 Results and discussion

5.1 SARIMA

5.1.1 Parameter identification

To construct a SARIMA$(p,\ d,\ q)(P,\ D,\ Q)_m$ model, we follow the Box-Jenkins procedure (Box and Jenkins, 1990) for time-series models. First, we establish that the time series being modelled is stationary using both DF-GLS, a version of the Dickey-Fuller test (a unit root hypothesis rejected at α=0.001, for 8 auto-selected lags), and the Kwiatkowski-Phillips-Schmidt-Shin test (a stationarity hypothesis cannot be rejected at α=0.1 for auto-selected lags). Thus, no differencing is required and we select the d parameter of the non-seasonal part to be 0.

Next, we identify the other two non-seasonal parameters using autocorrelation and partial autocorrelation plots (see Figure 2), as the number of lags at which the two functions enter the 95% confidence interval, thus suggesting p=1 and q=1. Examining ACF, we also find indications of seasonality: there are spikes at lags 7 and 8 and at 13 and 14 lags, but these spikes die down fairly quickly. This observation suggests a weekly seasonality ($m = 7$) as well as stationarity at the seasonal level.

Additionally, we tested different values for p and q as well as P and Q using Akaike Information Criterion, Bayesian Information Criterion, and Hannan-Quinn Information Criterion for time-series model selection. The results, shown

[2] http://www.gallup.com/poll/112723/gallup-daily-us-consumer-spending.aspx

[3] Available at https://nlp.stanford.edu/projects/glove/

Figure 1: Train-validation-test split in the CSI values.

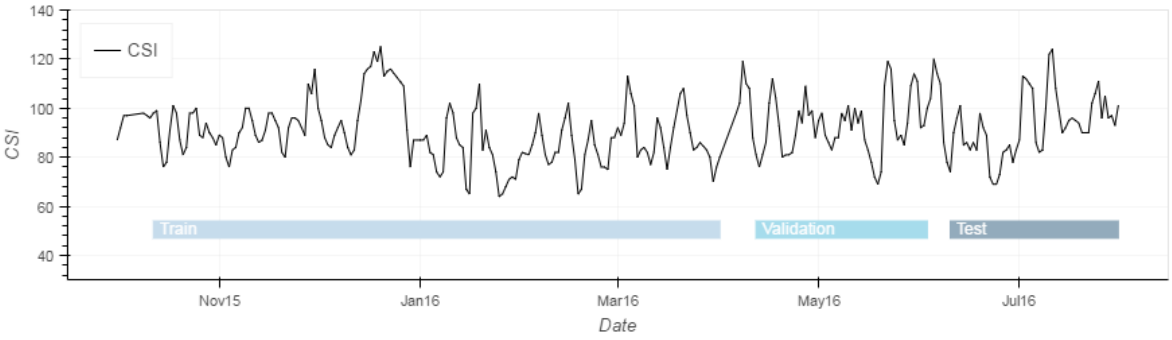
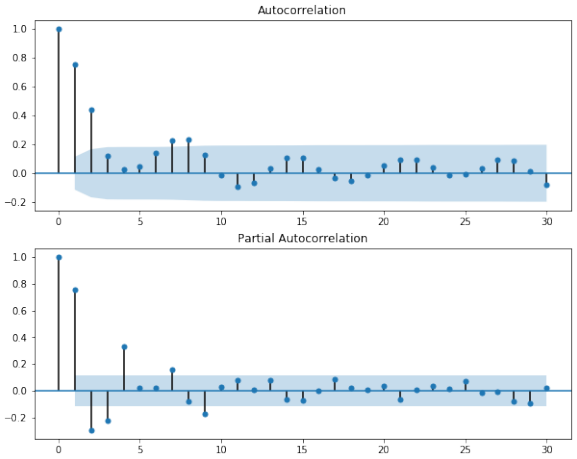


Figure 2: Auto-correlation and partial autocorrelation functions of CSI.



in Tables 1 and 2, largely agree with parameter identification based on ACF and PACF, and suggest that the optimal model takes the form SARIMA(1,0,2)(0,0,2)$_7$, which we thus used in further experiments.

AIC		BIC		HQIC	
(1,0,2)	**-267.49**	**(1,0,2)**	**-251.58**	**(1,0,2)**	**-261.03**
(3,0,4)	-267.22	(1,0,3)	-248.02	(1,0,3)	-259.37
(1,0,3)	-267.11	(2,0,2)	-246.76	(2,0,2)	-258.11
(1,0,4)	-266.48	(1,0,4)	-244.2	(1,0,4)	-257.44
(3,0,2)	-266.25	(3,0,2)	-243.98	(3,0,2)	-257.22

Table 1: Identification of non-seasonal AR and MA parameters in SARIMA based on Akaike IC, Bayesian IC and Hannan-Quinn IC.

AIC		BIC		HQIC	
(3,0,5)	-245.45	**(0,0,2)**	**-148.88**	**(0,0,2)**	**-203.74**
(0,0,2)	**-241.15**	(0,0,3)	-145.5	(0,0,3)	-202.24
(0,0,3)	-240.95	(1,0,2)	-144.99	(1,0,2)	-201.74
(1,0,2)	-240.45	(0,0,4)	-141.4	(3,0,5)	-200.29
(0,0,4)	-240.04	(2,0,2)	-140.63	(0,0,4)	-200.04

Table 2: Identification of seasonal AR and MA parameters in SARIMA based on Akaike IC, Bayesian IC and Hannan-Quinn IC.

	Horizon=1	Horizon=3	Horizon=7
Lag 0	14.93	13.72	13.74
Lag 1	**14.79**	14.16	**12.61**
Lag 2	15.85	14.57	14.91
Lag 3	14.88	14.37	14.72
Lag 4	16.02	16.28	15.97
Lag 5	15.03	**13.37**	14.07
Lag 6	15.21	14.78	14.7
Lag 7	15.27	14.86	14.12

Table 3: RMSE on the test set of SARIMA at different forecast horizons, for different lags between the day of registered purchase intentions and the forecasted CSI.

5.1.2 Lag length between purchase intention and spending index

Having identified the parameters of SARIMA for endogenous variables, we tested its quality with exogenous (i.e., semantic) variables supplied to it. To do that, we varied the number of days between the day of the CSI index and the day on which purchase intentions were registered that were used to forecast the index. These results are shown in Table 3.

The lag of one day seems a good choice: it is the best for the forecast horizons of 1 and 7 days, and

	Train	Validation	Test	$\Delta, \%$
Horizon=1				
Endogenous	7.15	13.46	15.09	–
Endog+Semantic	5.77	13.88	**14.79***	-1.9
Horizon=3				
Endogenous	7.15	13.83	15.44	–
Endog+Semantic	6.65	14.26	**13.37***	-13.4
Horizon=7				
Endogenous	7.15	13.46	15.09	–
Endog+Semantic	5.78	13.74	**12.52***	-17.03

Table 4: SARIMAX vs. baseline SARIMA. Improvements on the baseline are in bold, significant improvements (at $p < 0.05$) are indicated with an asterisk.

one of the best settings for the horizon of 3 days. It can be noted that for all the horizons RMSE values are considerably higher for lags greater than 1.

5.1.3 Adding exogenous variables

Table 4 compares SARIMAX with the optimal intention-index lag and the baseline SARIMA, for the three forecast horizons, on the train, validation and test datasets. The last column shows the difference of SARIMAX to the baseline as percentage of RMSE change. Statistical significance of the difference to the baseline was measured using a paired t-test. The results show that the addition of semantic variables leads to significantly improved forecasts, for all the three horizons, and the improvements tend to become greater as the forecast horizon increases: at h=7, the reduction in RMSE is 17%.

5.2 AdaBoost

5.2.1 Lag length between purchase intention and spending index

As the first step in experiments with AdaBoost, we examined different lags between the day on which purchase intentions were expressed and the day for which CSI was forecasted. To that end, we trained AdaBoost models on only semantic variables for different lag values. The performance of these models is shown in Table 5. Note that the results are the same for all the three forecast horizons, since the models included only on exogenous variables and past predicted values are not used to forecast the current value. These results suggest that the best lags are between 4 and 6 days, this contrasts with the findings for SARIMA, where the optimal was lag 1.

	AdaBoost	Gradient Boosting
Lag 0	15.34	12.91
Lag 1	14.78	12.98
Lag 2	14.38	**12.71**
Lag 3	14.24	13.59
Lag 4	13.59	13.18
Lag 5	13.8	13.09
Lag 6	**13.38**	12.86
Lag 7	15.21	12.86

Table 5: RMSE on the test set of AdaBoost and Gradient Boosting, for different lags between the day of registered purchase intentions and the forecasted CSI.

	Train	Validation	Test	$\Delta, \%$
Horizon=1				
Endogenous	5.94	8.86	9.61	–
Endog+Semantic	6.22	10.38	11.05	+14.9
Ensemble	7.16	10.16	11.92	+24.0
Horizon=3				
Endogenous	5.94	9.38	14.58	–
Endog+Semantic	7.39	12.69	**12.61**	-13.5
Ensemble	9.16	11.39	**14.51**	0.0
Horizon=7				
Endogenous	7.23	9.72	14.33	–
Endog+Semantic	4.73	11.99	**11.69***	-18.4
Ensemble	9.44	11.6	**11.94***	-16.6

Table 6: AdaBoost models with exogenous variables vs. Baseline AdaBoost.

5.2.2 Adding exogenous variables

Table 6 describes evaluation of two ways to introduce exogenous variables to forecast CSI with AdaBoost: the concatenation of endogenous and exogenous variables into one vector of predictors ("Endog+Semantic") and the ensemble method ("Ensemble", see Section 3.3). The last column shows each method's difference to the baseline ("Endogenous"). Because the experiments with SARIMA revealed that the CSI values have weekly seasonality, we use seven lagged values as endogenous variables in the AdaBoost algorithms. Exogenous variables are semantic variables at lag 6, which was found to be the optimal in the previous step.

Similar to the SARIMA results, these results also indicate that exogenous variables become beneficial as forecast horizons increase: at h=1, the baseline could not be beaten, but at h=3 and h=7 both methods which use exogenous variables

	Train	Validation	Test	$\Delta, \%$
Horizon=1				
Endogenous	4.51	9.48	9.52	–
Endog+Semantic	2.47	10.22	10.49	+10.1
Ensemble	6.46	9.04	**9.05**	-4.9
Horizon=3				
Endogenous	4.92	9.72	13.18	–
Endog+Semantic	4.55	10.83	13.28	0.0
Ensemble	8.56	9.43	**11.62**	-11.8
Horizon=7				
Endogenous	2.99	10.6	13.98	–
Endog+Semantic	4.55	10.68	**12.07**	-13.6
Ensemble	9.96	9.56	14.65	+4.7

Table 7: Best Gradient Boosting settings vs. Baseline.

improve on the baseline, often to a statistically significant level. The greatest improvement is achieved at h=7 with the concatenation method, which reduced RMSE by 18%.

5.3 Gradient Boosting

5.3.1 Lag length between purchase intention and spending index

As with the other regression methods, we first looked at the effect of the lag between the purchase intentions and the forecasted index on Gradient Boosting: for each lag between 0 and 7, a model was trained using only exogenous variables. The results are shown in Table 5.

While the best lag was found to be the lag of 2, the differences between the lags are not very prominent and tend to stay within 7% of each other. This result is still at odds with what was found for SARIMAX and AdaBoost. In subsequent experiments with Gradient Boosting, exogenous variables were used to forecast CSI at the lag of 2.

5.3.2 Adding exogenous variables

Table 7 describes the performance for Gradient Boosting models when exogenous variables are introduced via concatenation with the endogenous ones ("Endog+Semantic") and via an ensemble regressor that combines separate predictions made with endogenous and exogenous variables ("Ensemble"). The results again suggest that exogenous variables become helpful at longer forecast horizons: while at h=1 the concatenation method fails to outperform the baseline, and for the ensemble method the RMSE reduction is only 4.9%, the improvement on the baseline at h=3 and h=7 reaches 13.6%. The ensemble method tends to fare better than the concatenation method, but not consistently so: at h=7 its forecasts are worse than those of the baseline.

6 Conclusion

In this paper we have presented a new method to forecast consumer spending from purchase intentions found on social media, aiming to approximate responses of participants of traditional consumer surveys. In contrast to previous work that modelled economic confidence from the sentiment of social media posts, we use semantic models of nouns that are stated as intended purchases, which, on the one hand, helps to incorporate richer evidence available in the data, and on the other, creates low-complexity regression models. The utility of the data was evaluated using three popular forecasting methods: Seasonal ARIMA, AdaBoost, and Gradient Boosting regressors.

The key findings of this work can be summarized as follows. Adding information on intended purchases as exogenous variables alongside lagged values of the consumer spending index often yields statistically significant improvements over a baseline that is trained on the lag variables alone. The benefits are greater at longer forecast horizons: while we found little evidence of improvement at one-step ahead forecasts, at the horizons of three and seven days, exogenous variables reduced forecast errors by between 11% and 18% for all the regression methods. Furthermore, we analysed the optimal lag length between the day on which purchase intentions were registered and the day for which spending is forecasted, but could not find any lag values that would be consistently better than others across the regression methods.

As future work, we plan to further explore the proposed method on larger datasets. A particular interesting extension may be a comparison of this method to those that derive a prediction of consumer spending from search engine queries, considering that both approaches aim to capture consumer purchase intentions, but do so using very different types of user-generated content. Another promising extension may study techniques for eliminating the demographic bias present on social media, in order to create models that better approximate real-world data on consumer spending.

References

Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8):1485–1509.

Marco Baroni, Georgiana Dinu, and Germn Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 1:238–247.

George Edward Pelham Box and Gwilym Jenkins. 1990. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated.

Zhiyuan Chen, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Identifying intention posts in discussion forums. pages 1041–1050.

Kristof Coussement and Dirk Van den Poel. 2008. Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Inf. Manage.*, 45(3):164–174.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Piet Daas and Marco Puts. 2014. Social media sentiment and consumer confidence. In *Workshop on using Big Data for forecasting and statistics*.

Theologos Dergiades, Costas Milas, and Theodore Panagiotidis. 2015. Tweets, Google Trends, and sovereign spreads in the GIIPS. *Oxford Economic Papers*, 67(2):406.

Mohammed Elshendy, Andrea Fronzetti Colladon, Elisa Battistoni, and Peter A Gloor. 2017. Using four different online media sources to forecast the crude oil price. *Journal of Information Science*, 0(0):0165551517698298.

Yoav Freund and Robert E. Schapire. 1996. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Ifigeneia Georgoula, Demitrios Pournarakis, Christos Bilanakos, Dionisios N. Sotiropoulos, and George M. Giaglis. 2015. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. In *9th Mediterranean Conference on Information Systems*.

Mohamed Hamroun, Mohamed Salah Gouider, and Lamjed Ben Said. 2016. Large scale microblogging intentions analysis with pattern-based approach. *Procedia Comput. Sci.*, 96(C):1249–1257.

Stephen Hansen and Michael McMahon. 2016. Shocking language: Understanding the macroeconomic effects of central bank communication. In *NBER International Seminar on Macroeconomics 2015*. Journal of International Economics (Elsevier), Volume 99, Supplement 1.

Jane-Vivian Igboayaka. 2015. Using social media networks for measuring consumer confidence: Problems, issues and prospects. Master's thesis, University of Ottawa.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of 11th International AAAI Conference on Web and Social Media (ICWSM)*.

Brendan T. O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.

Samuel Rönnqvist and Peter Sarlin. 2015. Detect & describe: Deep learning of bank stress in the news. In *IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town, South Africa, December 7-10, 2015*, pages 890–897.

Steven L. Scott and Hal R. Varian. 2015. Bayesian Variable Selection for Nowcasting Economic Time Series. In *Economic Analysis of the Digital Economy*, NBER Chapters, pages 119–135. National Bureau of Economic Research, Inc.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Sinno Jialin Pan, Qing Li, and Huayi Li. 2014. Exploiting social relations and sentiment for stock prediction. In *Proceedings of EMNLP*.

Thársis T. P. Souza, Olga Kolchyna, Philip Treleaven, and Tomaso Aste. 2016. Twitter sentiment analysis applied to finance: A case study in the retail industry. In Gautam Mitra and Xiang Yu, editors, *Handbook of Sentiment Analysis in Finance*, chapter 23.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Web and Social Media*.

Simeon Vosen and Torsten Schmidt. 2011. Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, 30(6):565–578.

Lynn Wu and Erik Brynjolfsson. 2015. The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic Analysis of the Digital Economy*, pages 89–118. University of Chicago Press.