Purchase Intentions on Social Media as Predictors of Consumer Spending

Viktor Pekar
Operations and Information Management Dept., Business School
Aston University
Aston Street, Birmingham, B4 7ET, UK
v.pekar@aston.ac.uk

Abstract

The paper addresses the problem of forecasting consumer expenditure from social media data. Previous research of the topic exploited the intuition that search engine traffic reflects purchase intentions and constructed predictive models of consumer behaviour from search query volumes. In contrast, we derive predictors from explicit expressions of purchase intentions found in social media posts. Two types of predictors created from these expressions are explored: those based on word embeddings and those based on topical word clusters. We introduce a new clustering method, which takes into account temporal co-occurrence of words, in addition to their semantic similarity, in order to create predictors relevant to the forecasting problem. The predictors are evaluated against baselines that use only macroeconomic variables, and against models trained on search traffic data. Conducting experiments with three different regression methods on Facebook and Twitter data, we find that both word embeddings and word clusters help to reduce forecasting errors in comparison to purely macroeconomic models. In most experimental settings, the error reduction is statistically significant, and is comparable to error reduction achieved with search traffic variables.

Introduction

Forecasts of private consumption are an important tool used by governments and commercial organizations in many areas of their strategic decision-making. To build predictive models of consumer spending, researchers traditionally used a selection of macroeconomic variables, such as real personal income and interest rates on treasury bills, as well as measures of consumer confidence such as the University of Michigan Consumer Sentiment Index, see, e.g., Ludvigson (2004), Croushore (2005). The latter indicators are obtained via nationwide surveys, in which households are asked to comment on their expected economic situation.

In recent years, large search engine companies like Google and Baidu have opened access to current and historical data on the volumes of search queries submitted by their users. Because many queries imply a purchase intention, the data has been studied as possible evidence about future private consumption. Indeed, it has been found to be a useful leading indicator of demand in housing (Wu and Brynjolfsson 2015), automotive (Choi and Varian 2012), tourism (Li et al. 2017) sectors as well as the overall private spending in an economy (Vosen and Schmidt 2011; Kapetanios, Marcellino, and Papailias 2018; Woo and Owen 2019). As opposed to consumer confidence surveys, search engine queries are thought to capture concrete purchase intentions, and therefore are expected to more closely model consumer behaviour.

This paper explores purchase intentions expressed in the text of social media posts, as an alternative to search engine queries, in forecast models of private consumption. On social media networks, users describe their everyday lives, including their intentions to purchase a product or service, thus revealing information quite similar to the information available in search engine data. However, a search engine query is only assumed to indicate a purchase intention: in reality, searches for product names may relate to other information on the products, such as technical support. The text of a social media post, on the other hand, makes it possible to unambiguously identify purchase intentions using NLP tools.

Previous research developed different techniques to extract signals on economic indices from textual data, including counts of predefined keywords (Dergiades, Milas, and Panagiotidis 2015), topic models (Li, Shang, and Wang 2019), word embeddings (Rönnqvist and Sarlin 2015), sentiment analysis (Deng et al. 2018). In this paper, we study methods to construct semantic variables from social media posts to be used within forecasting models. To achieve that, we investigate word embeddings, which have proved to be an accurate representation of lexical meaning in many other NLP applications. We then propose clustering methods to aggregate words relating to purchase intentions into categories of goods and services, and use the categories as predictors in a forecasting model. A potential benefit of word clusters, as opposed to individual words or word embeddings, is that they can facilitate further analysis of factors impacting consumer demand. We introduce a new clustering method, which takes into account the temporal relatedness of the words to the consumer spending index, thus aiming to arrive at clusters of goods and services that are tailored to the problem of forecasting consumer demand. In the experimental part, we compare word embeddings and word clusters obtained from social media posts to Google Trends categories of queries and report on the relative usefulness of these types of predictors.

Related work

Our study is related to two streams of previous research, one of them being on models of consumer spending based on search engine data, and the other on methods to construct social and economic indicators from textual data.

Forecasting consumer spending from search engine data

Vossen and Schmidt (2011) use search traffic data on a broad set of Google Trends categories in an autoregressive model of private consumption in USA, and find they are better predictors than the popular Conference Board Consumer Confidence Index. Carriére-Swallow and Labbe (2013) build an ARMA model for automotive sales in Chile and then show that the introduction of an exogenous variable constructed from Google Trends leads to improved forecasts. Scott and Varian (2015) use Google Trends categories, among which the best predictors are selected using the Bayesian Structural Time Series procedure, in order to model the Consumer Sentiment Index by the University of Michigan. Wu and Brynjolfsson (2015) predict real-estate sales by incorporating search engine data into an AR model, along with other exogenous variables such as housing price index. Li et al. (2017) forecast demand in the tourism sector using Google Trends data. Kapetanios et al. (2018) develop a forecasting model of a retail trade index in three EU countries.

While these studies jointly point to clear usefulness of search traffic data for the consumer spending forecasts, such data has a number of disadvantages. A search query is assumed to reflect a purchase intention, but this assumption may be valid to some, but not all kinds of goods and services. Arguably the most useful type of search data are query categories, which summarize counts of semantically similar queries. The categorization is however task-independent, and the reasons for the particular categorization are opaque. For example, the Shopping category in Google Trends contains some subcategories relevant to consumer demand, such as Consumer Electronics and Apparel, but lacks other obvious ones such as Food and Beverages or Motor Vehicles, which are on the top level in the US BEA classification of personal consumption expenditure. Furthermore, one cannot find out which specific queries make up a category, and its meaning can only be judged from its label. These issues greatly limit the interpretability of models that use the search categories.

Extracting socio-economic indicators from text

Previous work has explored a variety of techniques to extract features from textual data that can be used in models of socio-economic phenomena. A popular predictor is based on sentiment analysis: the overall approach has been to analyze texts for sentiment, compile a daily sentiment index, use it as a variable in models of stock prices (Souza et al. 2016), currency rates (Georgoula et al. 2015), commodity prices (Elshendy et al. 2017), consumer confidence (O'Connor et al. 2010), or product sales (Cui et al. 2018). Sentiment analysis is known to be a hard problem in NLP, where high accuracy is difficult to achieve, especially without intensive domain or genre adaptation. Also sentiment does not always imply an intention: there is only a loose connection between actual consumer spending and sentiment found in posts on the topic of interest.

Other studies applied some form of lexical analysis of the posts in order to derive predictor features. These include using counts of predefined keywords (Dergiades, Milas, and Panagiotidis 2015) or hand-selected ngrams (Antenucci et al. 2014). To account for more complex semantics contained in the messages, a number of papers used their entire vocabulary in combination with a dimensionality reduction technique (Coussement and Van den Poel 2008), mapping the messages to semantic vectors, such as word embeddings (Rönnqvist and Sarlin 2015), or topic models (Hansen and McMahon 2016; Li, Shang, and Wang 2019).

The novelty of our approach is that, instead of using the full vocabulary of the posts, it detects phrases referring to goods and services that are stated as intended purchases, thus aiming to more precisely pinpoint signals about future consumer behaviour.

Methods

Macroeconomic predictors

We would like to assess if social media and search engine data provide useful evidence about future consumer spending, in addition to the information that is already available in macroeconomic variables. Thus, our baseline models include macroeconomic indicators that were used in a number of previous studies (Ludvigson 2004; Croushore 2005; Vosen and Schmidt 2011):

- Real personal income,
- Interest rates on 3-month Treasury bills,
- Stock prices (measured by the S&P 500 index).

Purchase intentions

The overall process we use to convert a collection of social media posts into time-varying signals predictive of a consumer spending index is the following.

In the first step, given a collection of posts, those containing purchase intentions are identified using lexico-syntactic patterns. The patterns are created from combinations of (1) first-person pronouns ("I", "we"), (2) verbs denoting intentions ("will", "'ll", "would like to", "want to", "wanna", "gonna", etc), and (3) verbs denoting purchase ("buy", "purchase", "shop for"). The text of posts that match the patterns is cleaned (emoticons, usernames, hashtags and URLs removed, most common "Internet speak" symbols replaced with regular words) and processed with a part-of-speech tagger. The head noun of the noun phrase following the purchase verb is then extracted (e.g., "headphones" in "I'd like

	Facebook	Twitter
Messages	79,046	589,137
Authors	74,308	466,998
Messages per author	1.06 (0.35)	1.26 (2.67)
Messages per day	199.1 (67.08)	1483.9 (399.8)

Table 1: The number of messages, unique authors, mean and standard deviation of messages per author and per day in the Facebook and Twitter datasets.

	K-Means	MajorClust	MajorClust-T
n	200	188	482
mean	4.85	5.16	2.01
st.dev.	3.77	18.94	6.0
min	1	1	1
max	18	219	122

Table 2: Descriptive statistics on sizes of clusters obtained with K-Means, MajorClust and MajorClust-T.

to buy new headphones"), and daily counts of the head nouns are recorded.

We represent extracted nouns as $\{n \in N\}$ and their counts over time T as $X_n = \{X_{n,t}: t \in T\}$. The consumer spending index, the target variable, is indicated by $Y = \{Y_t: t \in T\}$.

Word embeddings

A word embedding 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; after connection weights have been learned, the projection layer in the middle of the topology is taken to constitute a semantic vector for the word. The vector is a fixed-length, real-valued pattern of activations reaching the projection layer. Thus, on input, each word is represented as a co-occurrence matrix with a dimensionality equal to the vocabulary size of the training corpus (typically millions of words), and the method creates word representations of a much more compact size (typically several hundreds dimensions). The reduced dimensionality helps to reduce the complexity of the models, prevent overfitting, and is beneficial in computationally intensive classification and regression algorithms.

In our evaluation we include word embeddings, created with the word2vec method (Mikolov et al. 2013). For each date, we map each noun that was observed on that day to its word2vec vector that has been pre-trained on a large corpus of Twitter posts. The vectors of nouns registered for each day are then averaged to obtain a vector representing all purchase intentions expressed on that day. The components of the vectors will be used as variables in regression models.

	RMSE			**MAE**		
	train	val	test	train	val	test
AdaBoost						
AR	5.76	6.86	8.14	4.53	5.24	6.52
AR+ME	7.16	7.00	**7.88*****	5.42	5.46	**6.28*****
Random Forest						
AR	5.55	6.90	8.19	4.17	5.29	6.47
AR+ME	5.39	6.95	**7.90*****	4.00	5.39	**6.16*****
Lasso						
AR	6.89	6.45	8.08	5.19	5.03	6.38
AR+ME	6.90	6.54	**7.91**	5.19	5.11	**6.17**

Table 3: Error rates of models with (1) only autoregressive variables (AR) and (2) with autoregressive and macroeconomic variables (AR+ME).

Word clusters

To arrange the extracted nouns into categories, we experiment with two clustering methods: K-Means and MajorClust. Before clustering, the nouns are represented in terms of 200-dimensional word2vec vectors (Mikolov et al. 2013), pre-computed from a large corpus of Twitter posts available from the GloVe project[1].

K-Means. K-Means (Macqueen 1967) is one of the most popular clustering algorithms, well-known for its efficiency. Given a set of objects N represented as attribute vectors and an integer number k, the desired number of clusters, the algorithm searches for a partition of N into k non-hierarchical clusters that minimises the squared Euclidean distance between cluster members and the centroid of the cluster.

MajorClust. MajorClust (Stein and Meyer Zu Eissen 2002) is another non-hierarchical clustering method, but unlike K-Means, it does not require stopping criteria like the number of clusters to be pre-set in advance. The input to the algorithm is a NxN matrix of similarities (e.g., cosine) between objects in N. The algorithm begins by assigning every $n \in N$ to its own cluster. At each iteration, $n \in N$ gets reassigned to the cluster, to which it has the biggest similarity. An object's similarity to a cluster is calculated as the sum of its similarities to each of the cluster's members. Clustering stops when no object changes its cluster. MajorClust is known to benefit from cancellation of weak similarities from the similarity matrix. In this study, we keep 2.5% of pairs with the highest similarity values, tuning this parameter experimentally on the training set.

Both algorithms cluster nouns based on their general-domain semantic similarity, i.e. not taking into account semantic criteria that may be relevant for predicting consumption. For example, in the general case, nouns referring to certain kinds of accessories, watches and jewellery may be taken to belong to different categories, but in the context of models of private consumption, it makes sense to put them into a single category, Luxury Goods.

MajorClust-T. We introduce a modification of the Ma-

[1] https://nlp.stanford.edu/projects/glove/

Figure 1: Training-validation-test splits of CSI.

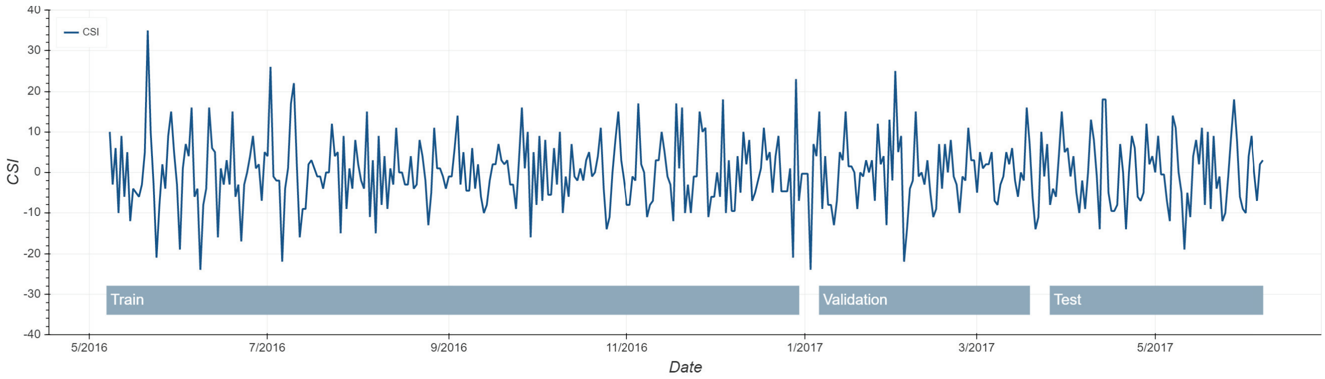
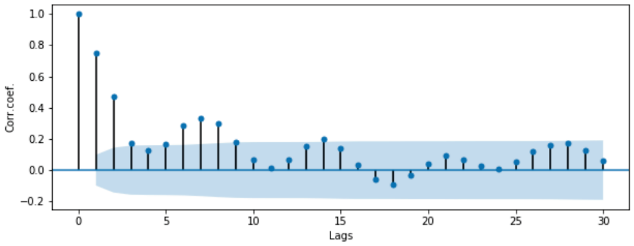


Figure 2: Autocorrelation plot of CSI.



jorClust algorithm, henceforth MajorClust-T, which aims to arrange nouns in clusters that are better suited for the task of modelling private consumption. We would like to encourage such groupings of nouns that are more predictive of the consumer spending index, and discourage those noun groupings that are less predictive of the index. To that end, we use the Granger causality test (Granger 1969), which examines if previous values of one variable are useful for predicting of following values of the other variable. This is achieved in the following steps:

1. In each pair of nouns $\{(n_i, n_j) \mid sim(n_i, n_j) > 0\}$ with a non-zero similarity in the input matrix, the training parts of X_i and X_j are selected and ensured to be stationary via first differencing (i.e., the differences between the current and the previous days' values were used instead of observed values) and the Augmented Dickey-Fuller test (Dickey and Fuller 1979).

2. Granger causality is tested between X_i and Y, and between X_j and Y, noting the likelihood ratio (LR) values of the tests.

3. Granger causality is tested between Y and the sum of X_i and X_j.

4. If the LR value of the test in step 3 is less than either of the LR values in step 2, the similarity value for n_i and n_j in the matrix is set to 0, in order to prevent their grouping during the actual clustering.

The modified similarity matrix is then input into the MajorClust algorithm to construct word clusters.

Because 2.5% of highest values are kept in the similarity matrix before Granger causality tests are applied, the time complexity of MajorClust-T has an overhead of only $0.025 \cdot \mathcal{O}^2$ in comparison to the original MajorClust algorithm.

Once clusters have been created, daily counts of cluster members are summed up to obtain daily counts of each cluster, which are then used as exogenous variables in a regression model.

Regression models

The general form of the regression model we use is as follows:

$$y_t = \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{p} \sum_{k=1}^{r} \omega_{j,k} X_{j,k} + e_t$$

where y_t is the consumer spending index at time t; y_{t-i} is the index lagged by i time steps; β_i is the coefficient of y_{t-i}; p is the maximum number of time lags; $X_{j,k}$ represents the count of k-th word cluster at time $t-j$; $\omega_{j,k}$ represents the coefficients of X_k at lag j; r is the total number of exogenous variables; and e is the error at t. The coefficients of the models are estimated by regression.

The textual data we use in our experiments is characterized by high dimensionality, and unlike much of prior work on forecasting consumer demand based on time-series models such as ARIMA (Vosen and Schmidt 2011; Wu and Brynjolfsson 2015), we select regression methods, that are capable of handling large amounts of predictor variables relative to the number of observations and are robust against noisy variables: Least Absolute Shrinkage and Selection Operator (Tibshirani 1994) and two ensemble decision tree regressors, AdaBoost (Freund and Schapire 1996) and Random Forest (Breiman 2001) algorithms.

	RMSE			MAE		
	train	val	test	train	val	test
AdaBoost	5.08	7.11	**7.63*****	4.05	5.15	**6.09*****
Random Forest	4.37	7.11	**7.75*****	3.19	5.32	6.17
Lasso	6.99	6.72	**7.78**	5.27	5.19	**6.11**

Table 4: Error rates of models with Google Trends variables added to AR and ME variables.

	RMSE			MAE		
	train	val	test	train	val	test
AdaBoost						
Word2Vec	5.34	7.14	**7.63*****	4.23	5.52	**6.03*****
MajorClust	5.12	6.85	**7.69*****	4.13	5.25	**6.13*****
MajorClust-T	4.01	6.97	**7.85****	3.50	5.39	6.32
K-Means	5.55	6.94	**7.74*****	4.33	5.32	**6.13*****
Random Forest						
Word2Vec	4.12	7.03	**7.51*****	2.90	5.45	**5.90*****
MajorClust	4.19	6.85	**7.73*****	2.96	5.15	**6.10**
MajorClust-T	4.06	6.96	**7.67*****	2.82	5.23	**6.06*****
K-Means	4.95	7.10	**7.87*****	3.68	5.42	6.21
Lasso						
Word2Vec	6.90	6.63	**7.67**	5.14	5.14	**6.11**
MajorClust	7.29	6.80	8.03	5.50	5.31	6.52
MajorClust-T	7.07	6.79	**7.82**	5.33	5.23	6.29
K-Means	6.89	6.82	8.02	5.15	5.36	6.48

Table 5: Error rates of models including Word2Vec, MajorClust, MajorClust-T and K-Means predictors derived from Facebook data.

Experiment design

Consumer Spending Index

As the target variable in our model, we use the Gallup Consumer Spending Index (CSI)[2]. The index represents the average dollar amount US households 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. In our study, we used the 3-day rolling averages of these amounts, spanning the period between May 7, 2016 and June 7, 2017, i.e. 397 days in total.

Social media posts

For the same time period, we collected public Facebook and Twitter posts that originate from the US and that express intentions to buy, following the procedure described above[3]. The sizes of the datasets, number of authors and frequencies of the messages are shown in Table 1.

	RMSE			MAE		
	train	val	test	train	val	test
AdaBoost						
Word2Vec	4.90	6.93	**7.67*****	3.93	5.43	**6.08*****
MajorClust	4.89	7.18	**7.68*****	4.02	5.50	**6.16*****
MajorClust-T	5.29	7.14	**7.61*****	4.21	5.45	**6.04*****
K-Means	5.48	7.06	**7.75*****	4.31	5.37	**6.14*****
Random Forest						
Word2Vec	3.88	7.03	**7.58*****	2.64	5.37	**5.91*****
MajorClust	3.92	7.15	**7.73*****	2.68	5.41	6.17
MajorClust-T	3.91	6.97	**7.68*****	2.68	5.35	**6.03*****
K-Means	3.93	7.07	**7.72*****	2.68	5.29	**6.04*****
Lasso						
Word2Vec	6.50	6.38	8.11	4.87	5.03	6.35
MajorClust	7.33	6.78	**7.83**	5.49	5.23	6.37
MajorClust-T	7.16	6.73	7.92	5.33	5.24	6.37
K-Means	6.85	6.83	7.94	5.14	5.38	6.38

Table 6: Error rates of models including Word2Vec, MajorClust, MajorClust-T and K-Means predictors derived from Twitter data.

From the collected messages, daily counts of nouns referring to purchases were extracted. To reduce noise from phrase extraction errors, we selected 1000 most common nouns which were then used to create word clusters.

The extracted nouns were used to create word embeddings vectors and word clusters using K-Means, MajorClust and MajorClust-T as described above. MajorClust came up with 188 clusters, and therefore for K-Means, we used k=200, for an easier comparison of informative features produced with the two methods. The statistics on the sizes of obtained clusters are shown in Table 2. One can see that K-means produces clusters of more uniform sizes, while the size of clusters in MajorClust and MajorClust-T are of much greater variance, and the largest clusters in them are several times larger than those of K-Means. Because many potential groupings of nouns were prevented in MajorClust-T, it produced more than twice the number of clusters of MajorClust and the mean size of clusters is much smaller.

Google Trends

The Google Trends (GT) website provides data on the volumes of queries to the Google search engine made since 2004, which can be searched by geographic and time criteria. Volumes of individual queries as well as hierarchical categories of queries are available.

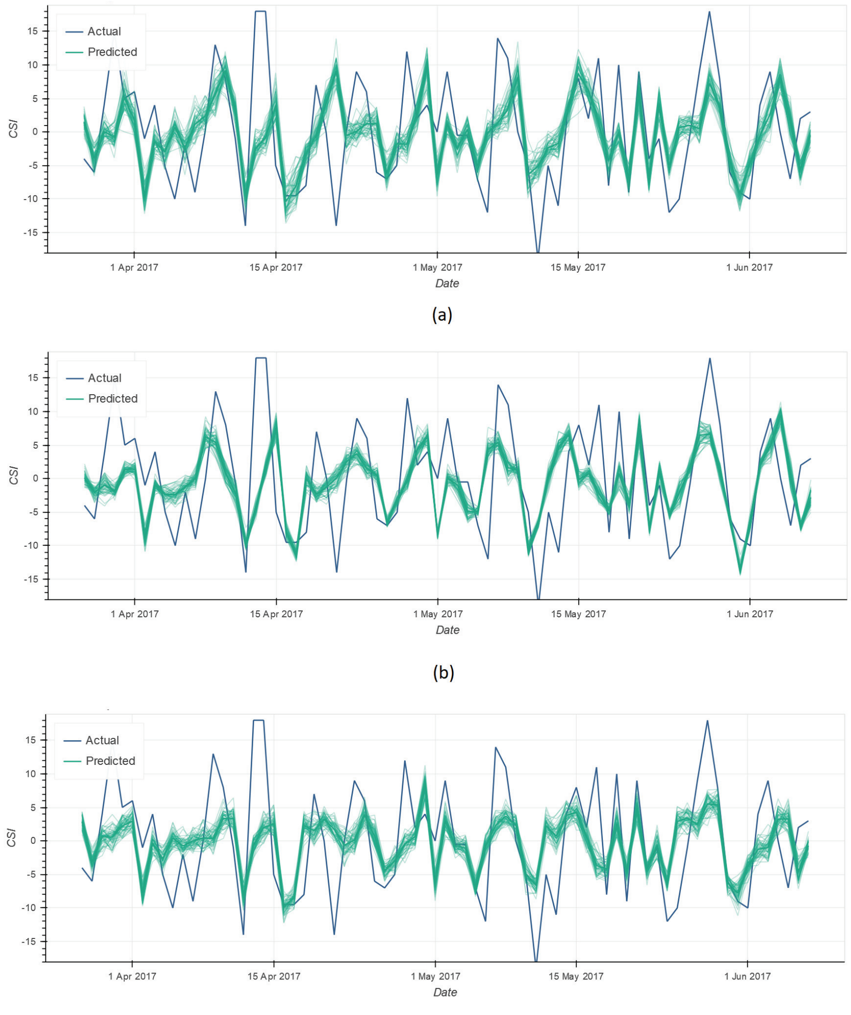
Following Vosen and Schmidt (2011), we select those GT categories that match the categories of the US Bureau of Economic Analysis classification of personal consumption expenditure[4], using the manual alignment of the two sets of categories developed in the original study by Vosen and Schmidt. For example, the BEA category ”Recreational goods and vehicles” is mapped to the GT categories

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

[3] The posts were retrieved via the Brandwatch Ltd. social media monitoring platform.

[4] https://www.icpsr.umich.edu/icpsrweb/NADAC/studies/36279

Figure 3: CSI values forecasted with Random Forests using (a) autoregressive and macroeconomic variables, (b) Google Trends variables in addition to the baseline variables, (c) word2vec variables in addition to the baseline variables.



(a)

(b)

(c)

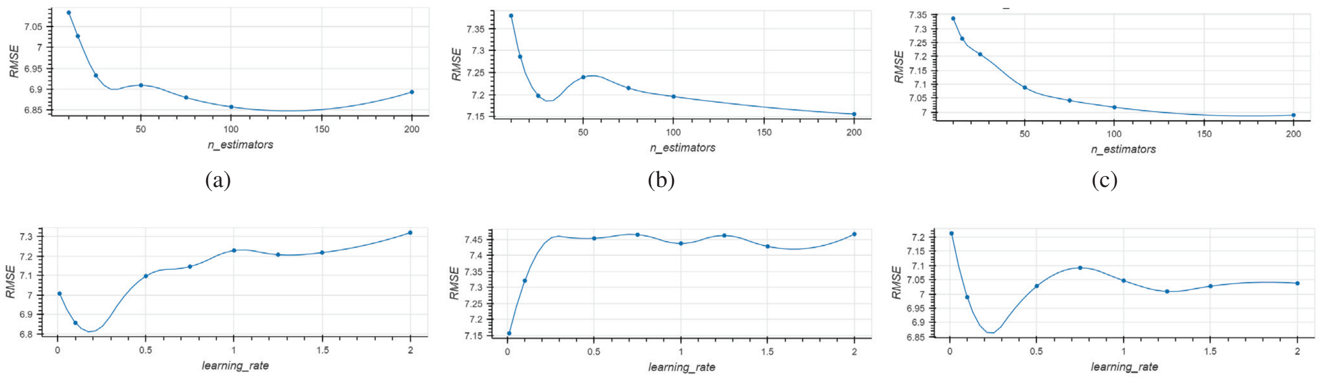
”Book Retailers”, ”Entertainment”, ”Entertainment Industry”, ”Movies”, ”Video Games”. In this way we obtained 51 GT categories (Vosen and Schmidt used 56 categories, five of these categories either had been since then removed from the GT categorization scheme or did not have any data in the relevant period), each of which was used as a predictor variable in the augmented regression models.

The volume of queries returned by GT is not the actual number of queries, but a normalized value, such that for any given retrieval criteria, the index is always between 0 and 100, 100 being the maximum volume among the retrieved datapoints. GT returns daily volumes for requests covering periods less than 6 months, and weekly volumes for periods greater than 6 months. Since the volume values are normalized relative to the maximum value in each specific request, we obtain daily query volumes for the entire 13 month period as follows. Weekly volumes for the entire period are retrieved, as well as daily data for all 5 months parts of the full set of dates. Then, within each part, we fit a linear regression on the weekly data, thus obtaining daily volumes for the entire period of interest.

Data preprocessing

Because days on which public holidays fell had no recorded CSI values, the missing values were supplied using linear interpolation. Further, CSI was found to be non-stationary according to the ADF (Dickey and Fuller 1979) and the KPSS tests (Kwiatkowski et al. 1992), and was therefore stationarized via differencing. The interpolated and differenced values of CSI are shown in Figure 1.

Examining the autocorrelation plot of CSI (see Figure 2), one can see that it exhibits weekly seasonality: there are significant correlations at lag 7 and further “humps” at lags 14, 21, and 28. This suggests that an autoregressive model of CSI should use 7 lags.



(a) (b) (c)

(d) (e) (f)

Figure 4: The number of estimators (top) and learning rate (bottom) of AdaBoost tuned on the validation set, for the models trained on AR+ME variables (a, d), on Google Trends (b, e), and on MajorClust-T clusters (c, f).

	RMSE			**MAE**		
	train	val	test	train	val	test
AdaBoost						
Word2Vec	5.49	7.17	**7.54*****	4.16	5.25	**5.99*****
MajorClust	5.21	7.11	**7.70**	4.09	5.16	**6.13**
MajorClust-T	5.01	7.12	**7.79**	4.08	5.28	**6.14**
K-Means	5.47	7.16	**7.60**	4.21	5.24	**6.06****
Random Forest						
Word2Vec	3.80	7.13	**7.56**	2.60	5.24	**5.99**
MajorClust	3.96	7.14	**7.68*****	2.76	5.22	**6.12**
MajorClust-T	4.44	7.18	**7.63*****	3.26	5.21	**6.02*****
K-Means	3.95	7.18	**7.70*****	2.77	5.26	**6.11****
Lasso						
Word2Vec	6.99	6.72	**7.78**	5.27	5.19	**6.11**
MajorClust	7.10	6.79	7.94	5.37	5.25	6.35
MajorClust-T	6.94	6.80	**7.78**	5.24	5.18	6.20
K-Means	6.76	6.72	7.92	5.06	5.23	6.24

Table 7: Error rates of models incorporating GT variables and four types of predictors derived from Facebook.

	RMSE			**MAE**		
	train	val	test	train	val	test
AdaBoost						
Word2Vec	5.33	7.13	**7.56**	4.11	5.22	**6.06**
MajorClust	5.44	7.17	**7.60**	4.22	5.25	**6.07**
MajorClust-T	5.22	7.17	7.89	4.14	5.42	**6.18**
K-Means	5.57	7.17	**7.56*****	4.30	5.23	**6.01*****
Random Forest						
Word2Vec	3.72	7.12	**7.64**	2.58	5.18	**6.10**
MajorClust	3.99	7.19	**7.76**	2.83	5.14	**6.12*****
MajorClust-T	3.74	7.12	**7.68**	2.56	5.18	**6.03**
K-Means	4.33	7.22	**7.62*****	3.22	5.24	**5.99*****
Lasso						
Word2Vec	6.93	6.72	**7.88**	5.22	5.21	6.21
MajorClust	7.16	6.82	**7.74**	5.37	5.26	**6.15**
MajorClust-T	7.03	6.75	**7.68**	5.26	5.25	6.17
K-Means	7.22	6.81	**7.78**	5.43	5.27	6.19

Table 8: Error rates for models incorporating GT variables and four types of predictors derived from Twitter.

Evaluation method

The available data was divided into the training, validation and test parts, in proportion 60%-20%-20%. 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, to ensure that no training data is used for validation or testing.

Once a model was trained on the training set and its parameters optimized on the validation set, it was applied to the test set to make one-step ahead forecasts. During training, feature selection was performed with Recursive Feature Elimination, determining the percentage of features to select by evaluating different amounts of the most informative features on the validation set.

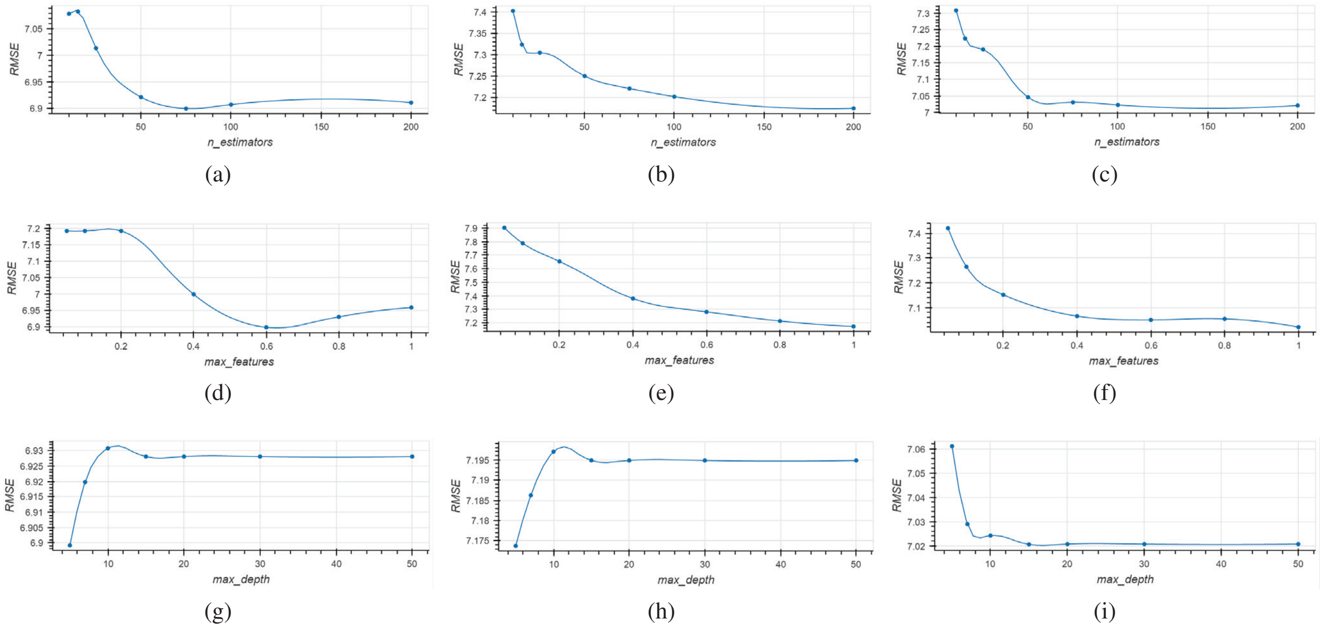
As evaluation metrics, we use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Both measure the differences between the model-predicted and ground-truth values, but RMSE gives greater emphasis to large, albeit rare errors than MAE, and so RMSE and MAE can be compared to detect presence of rare large errors.

During evaluation, the predictive power of the exogenous variables, i.e., word clusters and Google Trends categories, was assessed by measuring the extent to which they improve an autoregressive model in predicting CSI.

Results

The results reported below are the means of RMSE and MAE rates calculated over 50 runs of the same hyperparameter configuration of the regression method, with each run using a different random seed value, and thus starting from different initialization parameters. The reported significance of the differences in the mean scores was tested by the independent samples t-test for samples from populations with equal variances, and by the Welch test for samples from populations with unequal variances.

For a better perspective on the performance of the re-

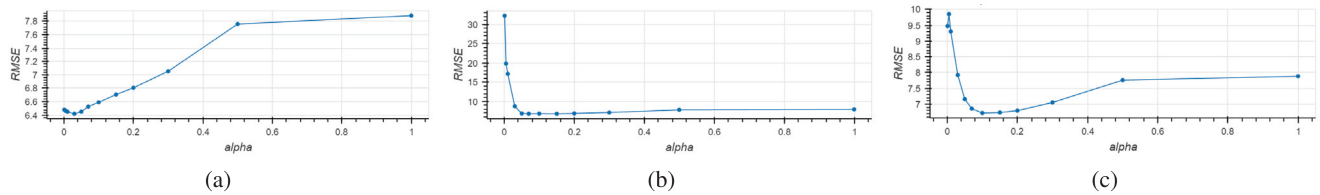


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5: The number of estimators (top), maximum features (middle) and maximum tree depth (bottom) of Random Forest tuned on the validation set, for the models trained on AR+ME variables (a, d, and g), on Google Trends (b, e, h), and on MajorClust-T clusters (c, f, i).



(a) (b) (c)

Figure 6: The alpha hyperparameter of Lasso tuned on the validation set, for the models trained on AR+ME variables (a), on Google Trends (b), and on MajorClust-T clusters (c).

	Variable	**Score**
1	movies	0.09
2	home-and-garden	0.08
3	restaurants	0.06
4	computers-and-electronics	0.03
5	automotive	0.03
6	health-insurance	0.02
7	auto-insurance	0.02
8	telecommunications	0.02
9	medical-facilities-and-services	0.02
10	electricity	0.02

Table 9: 10 most important Google Trends variables and their importance values.

gression methods, we included a persistence baseline, which simply output the preceding day's CSI value as the forecast for the following day. This method achieved the RMSE of 12.82 and the MAE of 10.18.

Baselines. Table 3 describes the performance of the baseline models[5]. The models built with only autoregressive variables are noticeably more accurate than the persistence baseline, indicating that AR predictors alone capture useful signals about future values of the target variable. Augmenting the AR model with the macroeconomic indicators improves forecast accuracy further for all the regressors (reduced error rates are shown in bold). The improvement corresponds to a drop in the error rates between 3.8% and 4.5% and is statistically significant for AdaBoost and Random Forests, in terms of both RMSE and MAE, at the $p < 0.001$ level. Thus henceforth, the AR+ME model will be used as the main baseline, against which other models will be com-

[5] Asterisks indicate significance at the *0.1, **0.05 and ***0.01 significance levels.

		K-Means		MajorClust		MajorClust-T	
Facebook	1	car, vehicle, truck, driver	0.07	car, vehicle, bike, truck, drive	0.08	book, world, record, story	0.08
	2	book, piece, movie, part, story	0.04	good, tonight, weekend	0.04	car, vehicle, bike, truck, suv	0.08
	3	couple, present, number, record	0.03	book, edition, hardback	0.03	iphone, nokia, device, tablet	0.04
	4	watch, tonight, weekend, night	0.02	food, beer, bottle, coffee, chip	0.03	food, beer, coffee, chip, pizza	0.04
	5	ingredient	0.02	game, xbox, warfare, cod	0.02	stuff, bunch, cause, alot	0.03
	6	beer, bottle, drink, water, wine	0.01	stock, pcs, price, import	0.02	game, player, play, ball	0.02
	7	coffee, milk, starbucks, mug	0.01	stuff, lot, bunch, other, amount	0.02	stock	0.02
	8	tank, gallon, air, tub, petrol	0.01	carrier, boat, flight, baggage	0.01	house	0.02
	9	tool, handset, electronics	0.01	gun, datum, weapon, van, mag	0.01	album, dvd, poster, soundtrack	0.01
	10	product, brand, model, unit	0.01	note	0.01	size, clothing, jacket, pant	0.01
Twitter	1	book, comic, manga, novel	0.05	food, pizza, dinner, chipotle	0.06	house, home, condo, mansion	0.06
	2	ticket, vip	0.04	ticket, shoe, shirt, clothes	0.04	food, vanilla, starbucks, beer	0.05
	3	product, supply, brand, oil	0.03	iphone, phone, xbox, laptop	0.04	tonight, weekend, summer	0.04
	4	bell, chipotle, taco, burrito	0.03	car, truck, bike, jeep, benz	0.04	dog, puppy, cat, kitten, pug	0.03
	5	shoe, pair, sock, cleat	0.02	tonight, good, weekend, work	0.03	iphone, phone, xbox, watch	0.03
	6	stuff, load, bunch, piece	0.02	beer, bottle, drink, glass, coke	0.03	makeup, brush, hair, color, skin	0.02
	7	music, pop, band, soul, rock	0.01	pop, rock	0.03	rover, tesla, mustang, cadillac	0.02
	8	lingerie, bikini, swimsuit	0.01	house, home, boat, insurance	0.02	edition	0.01
	9	rover, tesla, subaru, suv	0.01	dog, puppy, cat, bear, hedgehog	0.02	mom, mama, dad, daughter	0.01
	10	soda, coke, pepsi, redbull	0.01	ring, flower, necklace, candle	0.02	pill, balloon, chanel, vuitton	0.01

Table 10: 10 most important variables and their importance values in K-Means, MajorClust and MajorClust-T models trained on Facebook and Twitter data (for clusters having more than four members only the first four are shown.).

pared.

Google Trends. The forecast errors achieved by models incorporating autoregressive, macroeconomic as well as GT predictors, are shown in Table 4. Compared to the baseline, the addition of GT variables helps to significantly decrease RMSE for all the three regression methods, except MAE for Random Forest. The relative reduction of the errors is between 1.5% and 3%.

Facebook. Table 5 displays the forecast errors achieved by models that, in addition to AR and ME variables, included variables representing the semantics of phrases referring to purchase intentions in the Facebook data: Word2Vec, MajorClust, MajorClust-T and K-Means. We find that all the four types of semantic variables produce an improvement of RMSE on the respective baselines, at high significance levels. In terms of MAE, there is also an improvement for most of the semantic variables, except some of the clustering methods. The better performance in terms of RMSE suggests that the semantic variables are better at predicting large peaks and troughs of the target variable, but less useful for its smaller changes. Considering reduction of the error rates, the greatest decrease was achieved by the word embedding variables (2.6% RMSE and 3% MAE for AdaBoost, 5% RMSE and 4.6% MAE for Random Forest, 3% RMSE and 4.3 MAE for Lasso). Variables created by clustering nouns performed less well compared to word2vec, generally showing modest decreases in RMSE (between 1% and 2.5%), and either slight increases or only minor decreases in MAE (between 0.7% and 2.5%).

Twitter. Table 6 displays the results achieved with the same types of semantic variables constructed from the Twitter data. Similar to the results on the Facebook data, all these types of variables make it possible to reduce RMSE compared to the baseline at significant levels. We also find highly significant MAE reductions, with the exception of MajorClust variables. As with the Facebook data, the greatest relative reduction of the error rates is achieved by the word embeddings variables for the tree-based regressors (3.3% RMSE and 3.8% MAE for AdaBoost, 3.4% and 3.9% for Random Forest), although not for Lasso. The addition of cluster-based variables also consistently reduces the errors, but by smaller amounts (between 1.2% and 3.3%) and only for the tree-based regressors.

Comparing the results for MajorClust and MajorClust-T, we find that the use of temporal information for clustering leads to significant reductions in RMSE and MAE for Random Forests on both Facebook and Twitter data, and for AdaBoost on the Twitter dataset, at the 0.01 significance level. However, the reductions are rather slight (0.6%-0.9%). On the other hand, for AdaBoost on the Facebook data, we observe an increase in the error rates of 2%, also at the 0.01 significance level. For Lasso, MajorClust-T outperforms its counterpart by 1.5% on Facebook, but not on the Twitter dataset.

Combining social media and GT predictors. We next looked at whether forecasting accuracy can be further improved by combining GT variables with semantic variables created from social media data. Tables 7 and 8 report error rates achieved by the four resulting models for each regression method. The rates that are lower than the baseline are shown in bold; significant differences to both GT-only and semantics-only variables are indicated with asterisks.

We find that the combination of GT and social media variables consistently outperforms the AR+ME baseline. The best combination proves to be GT combined with K-Means variables: on both Facebook and Twitter data, for both types of regression methods, it reduces RMSE and MAE by up to 4.4%, the differences being highly significant for AdaBoost and Random Forest.

Considering the question whether the combination of the two kinds of variables improves accuracy compared to each kind being used on its own, the results are somewhat mixed: in some cases the combination helps to significantly reduce RMSE and MAE in comparison to either type of the variables, but in others the combination fails to deliver any further improvement. When error reduction was achieved, it is quite small: usually not more than 2.5%.

Figure 3 illustrates the test-set forecasting performance of the following Random Forest models: the baseline, i.e. the model incorporating autoregressive and macroeconomic variables (Figure 3a), the model which additionally uses GT variables (Figure 3b) and the model which uses Word2Vec predictors, in addition to the baseline variables (Figure 3c).

The plot for the baseline suggests that it tends to produce forecasts that (1) are often values very close to previous days' values: e.g., spikes in forecasts often follow spikes in actual values, and (2) miss spikes and troughs by large amounts: e.g., the spike just before April 15th and the second trough after the same date. The plots for the GT and Word2Vec variables, however, depict forecasts that better match the behaviour of the target variable: there are no forecasts that look like previous days' values predicted for following days, and there are fewer spikes and trough that are missed by large amounts.

Sensitivity of hyperparameters. We next examine hyperparameter sensitivity of the regressors trained on different types of predictors. Figures 4, 5, and 6 depict the values of several hyperparameters of AdaBoost, Random Forest, and Lasso, respectively, trained on the AR+ME variables, Google Trends categories, and MajorClust-T clusters. One can see that the algorithms trained on the GT and MCT variables show a much higher sensitivity for optimization. This is especially noticeable with Lasso: tuning Lasso's alpha parameter on the AR+ME variables results in the validation-set RMSE ranging between 6.4 and 7.8, whereas on the GT and MCT variables, different values of alpha cause RMSE to range between 6.7 and 33.0 (on GT) and between 6.7 and 9.9 (on MCT).

Informative variables. A potential benefit of cluster-based and GT models is that their variables can be directly related to categories of products and services, thus providing additional insights into which of them are useful leading indicators of consumer spending. Table 9 shows the most informative variables among GT variables. Table 10 shows the top 10 informative variables in models trained on K-Means, MajorClust (MC) and MajorClust-T (MCT) clusters from Facebook and Twitter data. The importance scores of the variables were calculated using the model-agnostic permutation feature importance algorithm (Fisher, Rudin, and Dominici 2018)[6].

We find quite a lot of lexically similar clusters across the clustering methods and datasets; there are also many similarities of the clusters to the most important GT categories. For example, in each list, there are clusters that have to do with:

- Automotive vehicles (e.g., in the Facebook datasets, the K-Means cluster ranked at 1, MC at 1, MCT at 2; in the Twitter datasets, the K-Means cluster at 9, MC at 4, MCT at 7; the GT variables ”Automotive” (rank 5) and ”Auto Insurance” (rank 7)),
- Food and beverages (e.g., K-Means ranks 6 and 7, MC rank 4, MCT rank 4, GT rank 3),
- Electronic goods (e.g., K-Means rank 9, MC rank 5, MCT rank 3, GT rank 4),
- Books and movies (e.g., K-Means rank 2, MC rank 3, MCT rank 1, GT rank 1).

[6] We use the implementation of the algorithm in the Skater library: https://datascienceinc.github.io/Skater/

The similarity of the most informative variables between the two datasets can be taken to reinforce the importance of these variables as leading indicators of consumer spending.

Discussion and conclusion

In this paper we aimed to establish if social media contains useful signals about a country's future consumer expenditure, beyond those available in macroeconomic variables that are commonly used for forecasting it. The study has investigated several methods to derive quantitative predictors from expressions of purchase intentions found in the text of public social media posts. The methods are based on detecting words referring to intended purchases and then constructing their semantic representations, based on word embeddings and on clustering words by their meaning. The clustering methods included the popular K-Means and MajorClust; in addition, we have proposed an extension of MajorClust that incorporates temporal information on word occurrence when building word clusters. The semantic predictors were evaluated by incorporating them alongside autoregressive and traditional macroeconomic variables into models of a consumer spending index. Furthermore, the study compared the effect of the semantic predictors on accuracy of the forecasts to the effect of predictors constructed from search engine data, which have been shown by recent research to be useful for forecasts of consumer spending.

Our findings can be summarized as follows. Predictors created from social media using either word embeddings or word clusters reduce forecasting errors in comparison to purely macroeconomic models. The error reduction is generally at statistically significant levels, and is on par with the reduction achieved with predictors constructed from search engine data. The best method to construct semantic predictors overall is word embedding, which consistently reduced the error rate by 2.6%-5% compared to the baseline, on both Facebook and Twitter data, for all the three regression methods included into the study. This level of error reduction, although not large in absolute terms, is similar to the reduction obtained by adding macroeconomic variables to the autoregressive model.

These results are in agreement with the studies by Asur and Huberman (2010) and Najafi and Miller (2015), who showed that purchase intentions expressed in social media help predict consumer demand. The positive impact of search traffic data on the forecasts is consistent with results of most previous studies (Vosen and Schmidt 2011; Scott and Varian 2015; Li et al. 2017; Woo and Owen 2019).

The introduction of temporal information into the MajorClust algorithm has helped to significantly decrease error rates for RandomForest regressors on both datasets, but AdaBoost and Lasso results have not conclusively shown the benefits of this information. As future work, a more detailed exploration the space of parameters of this method, possibly on multiple forecasting problems, will help to better reveal its strengths and weaknesses.

Furthermore, because both social media and search engine predictors were found to improve forecasts, but are likely to represent different kinds of purchase intentions, we looked at whether combining them in one model would improve performance of the models further. Here we obtain mixed results: in some experiments the combined sets of variables led to significant error reductions in comparison to both search engine and social media variables used on their own, whereas in others, we did not find any improvement of either one or both types of variables.

Our study has demonstrated the predictive power of purchase intentions in social media. Future work may focus on incorporating other types of information available in social media into the forecasting problem. The use of the structure of social networks seems a particularly promising avenue to explore. A number of previous studies have shown that the strength of social connections in a network can be operationalized in forecasting models of human behaviour (De Choudhury et al. 2013). Because purchase intentions, in particular, are known to be, to a considerable degree, guided by opinion leaders (Krauss et al. 2008), one would expect that this information can be also helpful in forecasts of consumer spending.

References

Antenucci, D.; Cafarella, M.; Levenstein, M.; Re, C.; and Shapiro, M. D. 2014. Using social media to measure labor market flows. Working Paper 20010, National Bureau of Economic Research.

Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 492–499.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Carrière-Swallow, Y., and Labbe, F. 2013. Nowcasting with google trends in an emerging market. *Journal of Forecasting* 32(4):289–298.

Choi, H., and Varian, H. 2012. Predicting the present with Google Trends. *Economic Record* 88:2–9.

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

Croushore, D. 2005. Do consumer-confi dence indexes help forecast consumer spending in real time? *North American Journal of Economics and Finance*.

Cui, R.; Gallino, S.; Moreno, A.; and Zhang, D. J. 2018. The operational value of social media information. *Production and Operations Management* 27(10):1749–1769.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Deng, S.; Huang, Z. J.; Sinha, A. P.; and Zhao, H. 2018. The interaction between microblog sentiment and stock returns: An empirical examination. *MIS Quarterly* 42(3).

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

Dickey, D. A., and Fuller, W. A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74(366a):427–431.

Elshendy, M.; Colladon, A. F.; Battistoni, E.; and Gloor, P. A. 2017. Using four different online media sources to forecast the crude oil price. *Journal of Information Science*.

Fisher, A.; Rudin, C.; and Dominici, F. 2018. Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective.

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

Georgoula, I.; Pournarakis, D.; Bilanakos, C.; Sotiropoulos, D. N.; and Giaglis, G. M. 2015. Using time-series and sentiment analysis to detect the determinants of bitcoin prices. In *9th Mediterranean Conference on Information Systems*.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424–438.

Hansen, S., and McMahon, M. 2016. Shocking language: Understanding the macroeconomic effects of central bank communication. In *NBER International Seminar on Macroeconomics 2015*. Elsevier.

Kapetanios, G.; Marcellino, M.; and Papailias, F. 2018. Empirical examples of using big internet data for macroeconomic nowcasting. In *CARMA 2018 - 2nd International Conference on Advanced Research Methods and Analytics*.

Krauss, J.; Nann, S.; Simon, D.; Gloor, P. A.; and Fischbach, K. 2008. Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS European Conference on Information Systems*.

Kwiatkowski, D.; Phillips, P.; Schmidt, P.; and Shin, Y. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 54(1):159–178.

Li, X.; Pan, B.; Law, R.; and Huang, X. 2017. Forecasting tourism demand with composite search index. *Tourism Management* 59:57 – 66.

Li, X.; Shang, W.; and Wang, S. 2019. Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting* 35(4):1548 – 1560.

Ludvigson, S. C. 2004. Consumer confidence and consumer spending. *Journal of Economic Perspectives* 18(2):29–50.

Macqueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Najafi, H., and Miller, D. 2015. Comparing analysis of social media content with traditional survey methods of predicting opening night box-office revenues for motion pictures. *Journal of Digital and Social Media Marketing* 3(3):262–278.

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

Rönnqvist, S., and Sarlin, P. 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*, 890–897.

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

Souza, T. T. P.; Kolchyna, O.; Treleaven, P.; and Aste, T. 2016. Twitter sentiment analysis applied to finance: A case study in the retail industry. In Mitra, G., and Yu, X., eds., *Handbook of Sentiment Analysis in Finance*. chapter 23.

Stein, B., and Meyer Zu Eissen, S. 2002. Document categorization with majorclust. In *Proc. 12th Workshop on Information Technology and Systems*, 1–6.

Tibshirani, R. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

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

Woo, J., and Owen, A. L. 2019. Forecasting private consumption with google trends data. *Journal of Forecasting* 38(2):81–91.

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