Supervised Learning Algorithms to Extract Market Sentiment: An Application using Commercial Real Estate Market

Steffen Heinig[a], Anupam Nanda [b]

[a] Liverpool John Moores Univeristy, Faculty of Engineering and Technology, Built Environment, Cherie Booth Building Byrom St, Liverpool, L3 3AF, United Kingdom
Email: s.heinig@ljmu.ac.uk | Tel:+44(0)151-904-1082

[b]The University of Manchester, Planning & Environmental Management, Manchester Urban Institute, 1.57 Humanities Bridgeford Street Building, Oxford Road, Manchester, M13 9PL, United Kingdom, Email: anupam.nanda@manchester.ac.uk | Tel:+44(0)161-306-8652

Abstract

Sentiment analysis has become a key area of research in economics and finance with methods evolving from traditional survey-based analysis to computational linguistic techniques. New developments in data handling and analysis have allowed extracting sentiment from vast amounts of written documents. However, these methods depend heavily on the existence of training and test data sets. The choice of training data is critical in such applications. We show a novel application from a unique market – commercial real estate. There are several unique attributes of the real estate market that makes such analysis critical for insightful market intelligence. In the absence of training data sets for the UK commercial real estate (CRE) market, we propose the use of Amazon book reviews for real estate related products. Our analysis has shown, that the use of more than 200,000 book reviews, can train different supervised learning algorithms, which in turn, can capture the sentiment and more importantly, it can help predict the direct commercial real estate market trends.

JEL Classification:

Keywords: supervised learning, machine learning, corpus, sentiment analysis, real estate

INTRODUCTION

The analysis of sentiment has become popular over the last few decades. In this paper, we provide a novel application using textual analysis and machine learning for sentiment analysis. We focus on the property market, which offers an excellent case for information asymmetry and is significantly influenced by the sentiment of the market players. However, a formidable challenge is the robust construction of a sentiment proxy. In the field of economics and finance, two types of measures are generally discussed, direct and indirect sentiment measures. Direct indicators are based on surveys or questionnaires with focused questions aimed at the target respondents. Whereas, indirect measures utilise macro-economic indicators that are related to the market of interest and may conceal elements of sentiment and statistical methods are required to extract the sentiment. Prominent examples of direct measures are the University of Michigan Consumer Sentiment Index, the Conference Board Consumer Confidence Index, the Survey of the Real Estate Research Corporation (RERC) or the Economic Sentiment Indicator (ESI). Several studies (such as Carroll *et al.*, 1994; Baker and Wurgler, 2007; Clayton *et al.*, 2009; Das *et al., 2015;* Marcato and Nanda, 2016; Heinig and Nanda, 2018; Heinig, Nanda, Tsolacos, 2020) have shown that sentiment plays a vital role in equity and real estate markets. The literature shows that surveys provide a better market sentiment than indirect measures. However, they should also be treated with caution. A group of interviewees influences the outcome of the surveys tremendously. They further require constant maintenance and the willingness of the interviewees to take part in the process frequently. The construction of survey-based measures can also be described as time-consuming.

Due to these shortcomings, a variety of indirect sentiment indicators have been developed or used, such as the use of online search queries (Choi and Varian, 2009; Preis *et al.*, 2010), the analysis of the trading behaviour of multi-asset property investors (Freybote and Seagraves, 2017) or the orthogonalisation of macroeconomic measures (Baker and Wurgler, 2006). However, indirect sentiment indicators do not measure the sentiment in the first place. Therefore, both direct and indirect measures might differ in their capability to predict the underlying market sentiment. The main problem when conventional sentiment proxies are used is the time difference between the measured sentiment and the publication date of the indicators. To generate the indicators, the proxy measures have to be published first. This generates a time lag, and uncertainty about the market arises. In fast-changing economic conditions, such time lag can render the indices outdated. More importantly, relying on such indices to devise policy interventions can be problematic as those will become time-inconsistent and not achieve optimal and desirable effectiveness.

In recent works, textual analysis has been used in understanding and measuring sentiment. The use of social media and mobile handheld devices have increased the amount of information stored in written text. At the same time, computing technology has become capable of processing large amounts of data. The analysis of text documents and the extraction of the market sentiment from them require large datasets over multiple years. Moreover, due to the presence of structured and unstructured information, uses of machine learning and its sub-category supervised learning have become popular. In this study, we combine textual analysis and machine learning and provide a robust application for the commercial real estate. Commercial real estate market provides an excellent application area as it is generally informationally inefficient, with infrequent transactions and spikes in investment volumes.

Supervised learning approaches are used to classify data entities based on an existing dataset. Different approaches can be used, such as support vector machines (SVM) or maximum entropy. In general, the algorithms will learn why one observation belongs to a specific category and not to another, and this process is also called pattern recognition. More precisely, a labelled dataset is split into two shares: one training and one test dataset. The training dataset is then used to teach algorithms the underlying pattern of the dataset. Since each entity is already classified, the algorithm mirrors the pattern in the dataset. After a validation process, the trained algorithms are used to classify the observations of the test dataset. Since the test dataset also incorporates the correct label, it is possible to judge how good the algorithms perform. After a satisfying level of prediction is reached, new and unlabelled observations can be incorporated in the classifiers. Since little is known about those items, the algorithms will classify each entity into one of the learned classes. This allows market participants to extract the underlying sentiment from a large dataset quickly and without actually reading any documents.

While various countries, such as the UK, offer direct sentiment measures (i.e. RICS survey), many countries don't. For foreign investors, this causes an investment hurdle. As the literature shows, direct sentiment measures are superior in comparison to indirect measures, and they are costly to produce and to maintain. Many scholars have tried to develop a substitute for those direct measures. Not only would that allow them to get suitable market insight, but it would also allow market participants to compare different markets with the same measure. Most of the proposed indirect sentiment measures rely on proxies based on other economic indicators, and those are country or market-specific. In the first case, these indicators are usually published weeks, if not months, after they are recorded. This means, that sentiment measures, based on these indices are likely to be outdated by the time they are published.

The generation of a labelled training corpus could be done manually, by actually reading and classifying the documents. Doing this on one's own would take much time. Splitting the workload among many people might cause the integration of biases. People are likely to have different opinions about specific topics. The judgement of the professional real estate market would need specialist knowledge. This is gained over a specific education and experience.

In this paper, we show how researchers and market participants can develop their sentiment classification system by adopting supervised learning methods. We tackle the task at hand from two sides. First, we identify the best algorithms based on different subsets of the training corpus; second, we use different dependent variables in the probit model to confirm the best textual sentiment measure. We use ca.150,000 newspaper articles over the cause of 2004-15 related to the commercial real estate market in the UK. Since no labelled real estate newspaper corpus is available, we recommend an innovative way to circumvent the issue. We use approx. 200,000 Amazon book reviews that are related to real estate products. We justify this approach, with the belief, that real estate related books are likely to be read by market participants and that they will use a similar jargon to the written text in newspaper articles. Using book reviews from Amazon provides the latitude, that a considerable corpus of written opinions is labelled. A supervised learning algorithm will adopt the underlying pattern in the reviews and classify the newspaper articles. Since the training corpus is of essential importance to the supervised learning algorithms, we use different subsets to train several algorithms. In a second step, we compare and evaluate our constructed textual sentiment measures to the Royal Institution of Charted Surveyors (RICS) direct sentiment measures for the commercial real estate market. Finally, we apply the selected sentiment indicators using standard models to see if they can predict the market.

Our results suggest that supervised learning algorithms can learn from Amazon book reviews. It is either the amount of training data or the applied sub-corpus, which increases predictability. We find further that a sentiment indicator based on newspaper articles is capable of reflecting both the market sentiment as well as direct economic measures.

We have organised the rest of the paper as follows. In the next section, we review relevant literature and situate our hypotheses within the literature. Then, we proceed to discuss our empirical framework and the data. Finally, we present the empirical analysis and perform some robustness checks, and conclude with a summary of key findings in the last section.

RELEVANT LITERATURE

The general sentiment literature divides between direct and indirect sentiment measures. This separation has also been respected in real estate. With regards to real estate sentiment measures, the literature has provided a series of different options. Publicly traded markets allow conclusions about the sentiment by utilising information about REITs. In Ling *et al.* (2014), eight different indirect sentiment proxies were used (i.e. REIT stock price premium to the Net Asset Value (NAV), the percentage of properties sold each quarter from the NCREIF index, the REIT share turnover, etc.). Private markets, on the other hand, require more farfetched sentiment proxies since the markets are not entirely dominated by professionals, here consumer spending and other macroeconomic factors play a crucial role. Private individuals have a different mindset as they trade their own homes they live in (Case and Shiller, 1989).

Both direct and indirect sentiment measures have been criticised by scholars for several reasons, but mainly because proxies do not measure sentiment in the first place, and surveys do not reflect the sentiment at the time when they are published.

More recent approaches allow for the quantification of text documents, as a new form of indirect sentiment measures. Newspaper articles, social media data or product/ movie reviews (He, 2012; Chen *et al.*, 2016), incorporate sentiment and opinions. Both scholars and market participants have identified these kinds of documents as a suitable source.

In the banking sector, for instance, textual analysis has been already applied for credit risk or asset valuation (Smales, 2016; Tsai *et al.*, 2016). Smales (2016) used the Thomson Reuters News Analytics tool for his analysis. A dataset which incorporates documents, which have been labelled by former market participants. The authors point out that a corpus, which has been annotated manually, generates much better results when the annotator has the background knowledge to the field discussed in the documents.

As already pointed out in the introduction, the use of supervised learning algorithms, which are likely to be used for the tasked at hand, requires the existence of training and test datasets. Unfortunately, no labelled newspaper articles corpus for the British commercial real estate market exists. Therefore, scholars either face the task to label a corpus manually or find a suitable way to bridge this issue. The process of labelling a text document has been in the centre of the discussion in various studies.

Following O'Keefe *et al.* (2013), a newspaper journalist tries to present the topic to a broader audience and is, therefore, addressing multiple opinions at once. Subsequently, this leads to a smoothing effect of the individual sentiments at the end. Based on the terminology of Liu (2012), the sentiment is usually expressed towards a topic. In this context, Liu (2012) stressed that opinion without a target is one without use. Based on this, Saif *et al.* (2016) and Lin *et al.* (2012) used a common sentiment topic method for their analysis. They identified that, within one text, multiple topics can be discussed and that the overall sentiment might differ from topic to topic. This increases the requirements for a topic-specific training dataset.

As each person processes information differently based on her education or social background, the manual annotation of text documents could be influenced by individual biases. In their study, O'Keefe *et al.* (2013) limited the number of annotators to three to guarantee consistency during the labelling process. They used the Fleiss kappa measure to illustrate how similar the results of the different annotators where. Chen *et al.* (2016) also underline that the annotation of a single user is worth more than the annotation of multiple users. This summarises the general issue when it comes to manual labelling of a text corpus and controls for the fact that only the social biases of one person influence the labels.

Another issue which scholars face in the absence of a labelled text corpus is that unfortunately, one could not recycle an off-topic existing corpus. For instance, Lin *et al.* (2012) state that labelled classifiers often fail to produce satisfying results within a new category.

While the development of textual sentiment indicators has been introduced to various other disciplines, applications to the real estate market are still sparse. Natural Language Processing (NLP) methods have been adopted in the equity market with success. Since real estate as an asset is not as frequently traded like stocks, researchers tend to apply equity market theories and new methods initially to the REIT market. Doran *et al.* (2010) have analysed the content of quarterly earnings conference calls of publicly-traded REITs and linked the tone of the calls back to the stock prices. They applied the proposed technique by Tetlock (2007/08) and used a customised dictionary and the *Harvard Psychosocial Dictionary*. Via the use of General Inquirer, the authors were able to extract the sentiment of the calls. Their analysis revealed that the Q&A part of those calls contributes more to the sentiment than the introductory speech of a chairman. A positive tone between the management and the analyst offsets negative feedbacks from negative company announcements. The authors were able to confirm the results for the equity market provided by Sadique and Veeraraghavan (2008).

Soo (2015) applied natural language-based techniques to the real estate market quite early. Motivated by the same observation as Case *et al.* (2012) or Foote *et al.* (2012), Soo (2015) thinks that the financial crisis has been analysed with a sole focus on the fundamental issues. The exclusion of sentiment and opinions is difficult to understand, given the behavioural finance knowledge to hand. The decision to focus on the housing market for her study is based on the fact that housing is more often traded by individuals and that sentiment shocks are more readily identified. The study examines all cities which are present in the Case-Shiller Home Price Index. Applying the method introduced by Tetlock (2007), Soo (2015) filtered the tone of the news articles to develop her underlying sentiment index. Similar to previous studies, she used the Harvard IV-4 Dictionary and included customised terms. Based on her study, she was able to forecast the financial market downturn with a lead of two years. The author showed that sentiment in news articles influences the real estate market.

Walker (2014a) extended the application of NLP to the real estate market. Based on a more significant corpus of news articles regarding the UK housing market, the author looked at the financial crisis and the influence of opinions which have led to irrational decisions. The results reveal that the sentiment or optimism in the market declined one year ahead of the crisis. Building upon those results, Walker (2016) showed that media coverage and influence on the behaviour of stock traders are much more far-reaching than assumed. He used news articles related to the UK housing market to see whether stock traders who trade UK housing company stocks are influenced by the sentiment of the articles. The results reveal that stock prices are influenced by the sentiment of the traders who are influenced by the sentiment of the housing market.

More recently, the application of NLP in real estate has been performed more dominantly. Heinig and Nanda (2018) have applied a classical bag of words approach. They tried to extract the sentiment from market reports from various UK based service agencies. Their results suggest that sentiment expressed in market reports mirrors the development of the market.

Hausler *et al.* (2018) applied SVM algorithms to the real estate market. They used newspaper article headlines. Their results suggest that headlines can foreshadow the property market. They used a labelled corpus of 5,000 headlines for the training process. Unfortunately, the authors did not provide any indication, how or who labelled those headlines. Assumed, that the authors did label them on their own, a measure of consistency, such as the Fleiss Kappa measure, could help to judge the quality. Besides, the author missed providing other quality measures such as the recall or the precision value for the trained algorithms.

This short review has revealed that the real estate market provides enough evidence for sentiment driven developments. The general separation into survey-based measures and proxy-based measures remain in the real estate literature, but the impression occurs that researchers use both measures in an interconnected way, when it is possible. It is striking that neither the literature nor the industry has been able to develop a general sentiment measure. However, due to the structure of the market and the different underlying interests of its players, it becomes clear that a generalisation of sentiment measures about entire markets and asset classes is nearly impossible.

For instance, surveys are limited to capture the entire market, by both the construction of the survey and by the target group, which is interviewed. Depending on the point of view of the interviewee, different sentiments can be assumed, and a private investor has a different sentiment when prices rise compared to a property vendor or a developer. It remains questionable if the sentiment of two opposing investor groups is the inverse function.

Nevertheless, this overview also shows that the application of NLP techniques and especially the use of supervised learning algorithms require well-fit training corpora. It is beyond doubt, that newspaper articles, which are linked to the real estate markets, provide enough sentiment to predict the market movement. Two aspects have become clear, first classifying a text document manually generates better results, than using a bag of words approach or any other machine-based classification method. Moreover, second, a classified corpus, which is trained on one specific topic, cannot be transferred to another unrelated topic.

However, labelling text documents manually are time-consuming and depending on the number of documents, also a monetary question. Services where one could hire people to label text documents, such as Amazons Mechanical Turk, would invite those people biases. Therefore, the results would be the same as the proposed method. Due to the absence of a classified training corpus, we suggest the use of book reviews for the training of three different supervised learning algorithms.

METHODOLOGY

It is necessary to point out that we are not the first ones who utilise Amazon product reviews for the sake of the extraction of sentiment. Several studies (e.g. He and Zhou, 2011; Zirn *et al.*, 2011, Min and Park, 2012; Reyes and Rosso, 2012; Moraes *et al.*, 2013) have identified the benefit of the reviews and the corresponding rating.

Focus is set on creating suitable proxies for the sentiment. In this study, we focus on Support Vector Machines (SVM), Maximum Entropy (MAXENT) and Random Forest (RF) algorithm.

Based on the literature *SVM* has been used widely for the classification of text documents [Bai (2011), Chen C. C. *et al.* (2011), Fan *et al.* (2011), Walker M. A. *et al.* (2012)]. Nguyen *et al.* (2015) state that *SVM* can handle high dimensional data, which is a good reason why the algorithm is very competitive when it comes to text classification. Medhat *et al.* (2013) also state that *SVM* is a suitable method for text documents since the sparsity of text allows for a linear classification since the features themselves are irrelevant but tend to correlate. *SVM* belongs to the class of linear classifiers.

In general, the method tries to find the best linear separation between the different classes. The linear separator is called a hyperplane. Initially, *SVM* was applied to binary classification problems, where a linear separation only needed to be achieved between two categories. According to Cortes and Vapnik (1995), the method in its simplicity is based on the assumption that there is a vector $\overline{w}$ of any length which is perpendicular to the median line of the hyperplane and vector $\overline{u}$ which is an unknown data point.

However, the issue with text data is that it more likely resembles a multiclass issue. At the same time, too many categories are likely to cause issues during the classification process. The original idea of classifying the news

articles based on the star system of *Amazon* (five categories) has not produced any satisfying results.[1] The reasons for this might be that the calculation of this number of options has reached its limits. However, the reduction of classes to three (positive-neutral-negative) has produced results.[2] In the literature, the classification of text into more than two categories is described as a multiclass classification issue. The proposed approaches are *one-versus-all* and *one-versus-one*. Hsu and Lin (2002) state that the *one-versus-all* approach calculates *n SVM* models, where *n* represents the number of classes, and then decides for each data point when a maximisation has been realised. This assignment is based on probability. This process is computationally expensive since multiple data points are calculated at once for multiple models.

Another classifier is the maximum entropy classifier which belongs to the class of probabilistic classifiers. A reason for the use of this distribution is that it is uniform. Uniformity equals higher entropy which is desired in this context since no pre-knowledge of the dataset is assumed. A MAXENT classifier quantifies the uncertainty of the dataset. It is expected that the distribution maximises the entropy by minimising the commitment and that it should be similar to some training data. Therefore, some constraints are introduced. The approach allows for different specifications, which are based on the data and our expectations. In a case where no constraints are introduced, the classifier assigns to each event the same probability. If there is pre-knowledge of the data and its distribution, then we could assign different expected distributions to each micro-stage. To summarise, the best model created by a *MAXENT* classifier is the one which allows for the most uncertainty from the data.

Similar to a BAGGING approach, where decision *TREE*s are used for the classification problem, the *RANDOM FOREST* does also rely on this method. Introduced by Breiman (2001), the approach adds more randomness to the construction of *TREEs*. In general, the nodes of the *TREE*s are split among all variables. In a *RANDOM FOREST* approach, these nodes are split based on the best of a subset of predictors, which are randomly chosen at each node [Liaw and Wiener (2002)]. Multiple *TREE*s are grown at the same time, and then the best predictor for each subset is selected by vote. The two essential measures for the *RANDOM FOREST* approach are the accuracy of the classifiers and the identification of how independent they are (correlation). *RANDOM FOREST* approaches can also be modified with different kernel parameters, which will improve the overall performance of the classifiers.

Once we have created suitable sentiment proxies using the above methods, we apply this within a traditional discrete choice model, such as Probit models, to detect changes within the underlying market. The calculation of the referring probabilities and the application of this model class has been widely used in real estate (see Tsolacos, 2012). The dependent variable in probit models is dichotomous and takes the values 0 or 1. We have decided to use the change of the *MSCI* all property growth rate for all assets and offices ($MSCI$). The two dependent variables are available monthly from January 2004 to February 2017, with a total of 158 observations.

$$\Pr[MSCI_t = 1] = \Phi\left(\sum_i f(textSent_{t-i})\right) \tag{1}$$

with $MSCI_t = 1$ if the monthly overall growth rate is negative at time t and vice versa. The different textual sentiment indicators $f(textSent_{t-i})$ are applied to the model, with the later in this study to determine the lag structure, via the use of the AIC.

We will not apply all constructed indicators, but those which have been proven statistically relevant. Pr is the probability forecast for the dependent variable at time t, given the cumulative density function of the normal distribution.

Equations 2 and 3 states the empirical models,

$$\Pr[MSCI\_cg\_aa\_ap_t = 1] = \alpha + \sum \beta_i \, textSent_{t-i} + \varepsilon_t \tag{2}$$

$$\Pr[MSCI\_cg\_aa\_omtwe_t = 1] = \alpha + \sum_i \beta_i \, textSent_{t-i} + \varepsilon_t \tag{3}$$

[1] We stopped the calculation after more than 48 hours, or in other cases, the calculation was automatically stopped by the program. The calculation was performed on two different computers: 8GB and 128GB RAM machines.

[2] The R package [e1071 by Meyer at al. (2014)] does offer for *SVM* the specification of kernel parameters. In this study, we have not applied any specifications, and the model has produced results for the three categories. There might be a possibility that the results could be improved by specific kernel arguments.

with α and β_i being coefficients, which will be estimated. ε_t refers to the normally distributed error term. The textual sentiment represented by ($textSent_{t-i}$). The dependent variables, as dichotomous growth rates for all assets and all properties ($MSCI\_cg\_aa\_ap_t$), for all offices ($MSCI\_cg\_aa\_ao_t$), for all offices in the City (London) ($MSCI\_cg\_aa\_oc_t$) and all offices in Mid Town and West End (London) ($MSCI\_cg\_aa\_omtwe_t$).

DATA DESCRIPTION

We use three different datasets in this study. For the training of the three supervised learning algorithms (Support Vector Machines, Maximum Entropy and Random Forrest), we are using Amazon product reviews. Newspaper articles are used for the extraction of the market sentiment. For evaluation of the constructed sentiment indicators, we utilise the MSCI IPD property series for all properties and all offices.

Amazon Data (training data)

The first dataset of this study consists of *Amazon* real estate related book reviews. We have crawled over 224,000 book reviews from around 5,800 different products (mainly books) from www.amazon.co.uk.[3] Each book review has a rating between one (negative) and five stars (positive). The books were selected by the following search terms:

> *real estate investment, property investment, real estate economics, real estate finance, real estate private equity, real estate valuation, property management, property valuation, property finance and real estate investment trust.*

Taking a closer look at the data, two things become clear. The crawling process downloaded a range of reviews for books which are not related to real estate (e.g. intellectual property) and second, people tend to rate the books more positively. In the collected dataset, 57% of all reviews are rated with five stars.

This creates another issue for the labelling process. A model that is trained on the raw data would favour the neutral or positive category. We have, therefore created five different datasets to control for these biases. While training corpus one is using the Amazon data unchanged, a smaller training dataset (corpus 3), is equally distributed over the five categories with a total of 37,740 reviews (7,548 reviews per category)[4].

The literature suggested the use of three (positive, neutral and negative) rather than five categories. We have created, based on the initial corpora, another three training corpora with just three sorting options (corpus 2, 4 and 5). Over the training and testing process, the machine learning algorithms seem to perform better when they encounter fewer sorting options. Again, corpus 2 is using the initial dataset, where we have just aggregated the two bad (1 and 2 stars) and good (4 and 5 stars) categories. A similar approach was taken for the construction of training corpus 4. Based on the equalised five-category corpus, we aggregated category 1 and 2 as well as 4 and 5, however, left the neutral category unchanged with 7,548 observations. Finally, training corpus 5 is just using three categories from the initial dataset. We have used 10,221 reviews for one, three and five-star rating. Transforming the star ratings into the categorical ratings leads to a shift in the categories. One and two stars are transformed into negatives, three stars become neutral, and the remaining two have been assigned to the positive category.

The newly assigned categories have shifted more weight to the negative and positive categories in the equal training corpus and much more weight to the positive category in the training corpus, which uses all reviews. The last issue is around labelling. On a linguistic and subjective level, some of the given ratings seem out of order. However, we wanted to interfere as little as possible in this initial trial. Yet, it seems debatable that "ok" as a stand-alone comment has a rating range from 1 to 5. The same applies to "awesome" or "excellent": subjectively we would rate books with these comments on the upper scale.

[3] The website was accessed on 12 March 2018.
[4] 7,548 did represent the lowest number of observations for the 2-star rating.

Newspaper articles

The main dataset has been collected via ProQuest UK News & Newspapers. The service provided access to a variety of UK based newspapers and was formerly known as UK Newsstand.

We performed a search on a monthly basis as the website only displays approximately 1,000 articles per search. The search function of the tool, which allows the pre-filtering of articles, is highly sensitive to the search terms. The data were collected with the following parameters:

> *English language, newspapers in the UK and full-text search; and with these search terms: Savills, BNPPRE, DTZ, Jones Lang LaSalle, JLL, Cushman & Wakefield, office property, retail property, commercial property market, REIT, real estate investment trust and London.*

A total of 118,842 articles were displayed. However, during the crawling process, only 109,103 articles were downloaded. Reasons for this are unknown. Each entity is identifiable by date, publisher, title and full text of the article.

Even though the search terms aimed to be focused on the real estate market, this original corpus seems to be noisy. We have therefore decided to construct several sub-corpora, which in our opinion, reduce the noise within the corpus. This follows the idea of other researchers that the sentiment should be analysed towards a specific feature [Liu (2012)]. The search parameters also collected several housing-related articles; therefore, the first sub-corpus excludes all housing articles. We removed all articles which included the words:

> *residential, housing, home, apartment or house;*

this reduced the number of articles from 109,103 to 62,266. However, this general exclusion might have excluded articles which discussed the broader real estate market. Nevertheless, we assume that the smaller corpus does focus more on the commercial real estate market.

A second sub-corpus was created and only includes articles with the word *London* (74,266 articles). That does not mean that all articles solely analyse the London real estate market; however, the chances are high that the property market of the city is at the centre of the discussion.

We are further interested in whether newspapers with a circulation above 100,000 papers per day might be able to influence the market more deeply; so, the third sub-corpus only includes:

> *The Daily Mail, the Daily Record, The Evening Standard, The Financial Times, The Daily Mirror, The Daily Telegraph, The Guardian, The Sun and The Times (52,954 articles).*

Since we would like to examine the commercial real estate market and how market participants are influenced by news, we have further decided to look only at Financial Times (FT) articles with the assumption that real estate finance professionals are more likely to read the FT than other newspapers (11,948 articles).

[insert Table 1 here]

Table 2 compares the two datasets with each other. Looking at the sparsity matrix is becomes apparent that for the full training dataset, a total of 224,395 reviews were collected. However, only 580 different terms are included in the corpus. The longest term has 13 characters. 97 % of the matrix is sparse, meaning that 126,410,604 cells of the document term matrix (DTM) are empty. The DTM is created by the number of documents and the total number of terms in the document. Different to the Amazon results, we see that by half of the documents nearly 9 times as much terms have been used. This seems logical, given the characteristic of both document types.

[insert Table 2 here]

MSCI Data

For the probit model, where we will test whether the textual sentiment indicators can predict the CRE market, the *MSCI* all property all asset, all office, office in the city and offices in Mid Town and West End capital growth indices will be used (Table 3:7). All will be modified into a binary or dichotomous variable with values of 0 and 1. One will represent those instances with negative growth.

The *MSCI* data is available on a monthly level from January 2004 to December 2015, which provides a total of 144 observations. According to the IPD Index Guide, "capital growth is calculated as the change in capital value, less any capital expenditure incurred, expressed as a percentage of capital employed over the period concerned".

Since no transactions, within the index-construction period, are considered[5], all series are essentially valuation driven. The index should only reflect the actual market returns and should ignore unusual developments of the property, which are caused by the individual management. We are aware of the question, whether the chosen dependent variable is suitable or not. It remains unclear if the reaction of the market or the reaction of the appraisers is measured. We assume that there is a fair chance that the blurring of multiple valuations, performed by different valuers should overcome this issue. Each valuation is based on assumptions taken from the market. These assumptions should be corrected or at least updated given new developments within the market.

[insert Table 3 here]

RESULTS

Unreported results have confirmed that our initial thoughts were correct. We did analyse the various training corpora with the help of a performance test. Here roughly 20% of the training data was prior removed before the algorithms were trained. These new algorithms were then tested against this withheld data to check their performance. As it turns out, corpus 5 has produced the best performance in these tests for all three different approaches (SVM, MAXENT and RANDOM FOREST). We have, therefore, decided to use these algorithms for the remainder of this analysis.[6]

We have used these three algorithms to create a set of 15 textual sentiment indicators—five for each algorithm with one for each specific sub-corpora. In the next step, we perform a correlation analysis between the constructed sentiment measures and the RICS market surveys. Finally, we use a simple probit approach to justify the use of the selected algorithms.

Correlation analysis

Since little is known about the quality of our constructed textual sentiment indicators with the regards to the commercial real estate market, we now like to test if the constructed measures show any relationship to the directed measures for British CRE market. Several studies (see literature review) have shown that direct measures perform better in comparison. Therefore, a moderate to a strong correlation between the two different types could confirm that the constructed measures can pick up the CRE market sentiment.

[insert Table 4 here]

Table 4 illustrates the correlation between direct and indirect sentiment measures. It can be seen that those measures which are based on the full corpus performed best—looking at the overall RICS survey we see that for the SVM algorithm the full corpus (0.575) and the London based corpus (0.558) produced strong correlations. For the MAXENT measures (0.635) only the full corpus measure reached a correlation above 0.5. The RF approach instead produced for the full corpus and the no-housing corpus strong correlations with the RICS direct measure. Those results are slightly improved by the office sales and rent survey measure. Here the highest correlation was achieved by the MAXENT full corpus measure, with a correlation of 0.66. For the retail-based measure only weak to moderate results were achieved. Only the London based RF algorithm produced a correlation above 0.5.

These results are surprising, as we believed that the focused sub-corpus would perform better. So the no-housing or the London specific corpus sentiment index should have resulted in a higher correlation with the direct measures. However, we are now able to confirm that our constructed measures can extract, at least to some extent, some of the underlying market sentiment.

[5], please refer to https://www.msci.com/documents/1296102/1378010/Indexes+and+Benchmark+Methodology+Guide.pdf/bfbd2637-581d-411e-bd5f-34d0d2b6b9c1, accessed on 22.11.2018

[6] These performance test results and all explanations are kept out of the main body of the paper due to brevity and those are available upon request.

Probit Model Results

In this analysis and due to the previously shown results, we test the three textual sentiment measures, which are based on the full newspaper corpus. They have revealed the best correlation with the direct RICS sentiment measures. Using a simple probit model, we examine how well the indicators can reflect the actual CRE market development. Due to the advantage of the constructed sentiment measures, we can switch to a monthly analysis since the four different MSCI market measures are available on a monthly base.

Table 5 illustrates the results for the first probit model, using the MSCI converted capital growth rate for all assets all properties. All three textual sentiment measures have a negative impact on the dependent variable and are significant at a 5% and 1% (MAXENT) level, respectively. The chosen lag structure was estimated by lowering the Akaike Information Criteria (AIK). Different from other indirect sentiment measures, it can be seen that the maximum number of lags is rather small, not only in this probit model but in the other three as well. Despite the significance of the indicators they perform quite weakly, the highest pseudo-r-square value was reached by MAXENT classifier with 0.048. Therefore, we find a rather weak result of the classification analysis and the analysis of the area under the Receiver Operating Characteristic (ROC) curve.

[insert Table 5 here]

The second probit model (Table 6) reveals the best results, both in terms of the pseudo-r-square and the classifications. All three indicators show a negative sign and are highly significant. The highest pseudo-r-square was reached by the MAXENT model (0.334), which also produced the best overall classification (88.19) and the highest value for the ROC curve (0.87).

[insert Table 6 here]

CONCLUSION

Sentiment plays a very important role in the economic decision-making process. The root cause of this role of sentiment lies with the fact that the world of economics is uncertain, full of asymmetric information. In most economic transactions, information about the product, process and future value does not flow seamlessly among the relevant stakeholders. Moreover, different stakeholders hold a varying level of quantity and quality of information. This asymmetry makes economic agents form an expectation and take risks under uncertainty. The belief and conviction of economic climate are what we can call as 'sentiment'. A clear positive or negative sentiment can lead to a transaction decision much quicker than the neutral sentiment or weak sentiment. Many researchers have attempted to extract such sentiment information hidden in economic variables. There are two notable failures in this regard – lack of relevant data on sentiment and methodological constraints. In this paper, we focused on the language of sentiment in a sector where significant information asymmetry exists. The business of real estate is fraught with information asymmetry, which makes an understanding of sentiment as key to analysts, policymakers and market players for investment decisions and policy formulation.

We have put together a new application area (i.e. real estate sentiment) with supervised learning methods. Our constructed measures are able to extract, at least to some extent, the underlying market sentiment. The results do indicate superiority over traditional methods. Our results suggest that supervised learning algorithms can learn from Amazon book reviews to a large extent. The level of learning depends on the amount of training data and applied sub-corpus. Both seem to add to predictability. We also find that a sentiment indicator based on newspaper articles can reflect both the market sentiment and economic indicators.

We like to highlight that we have used the entire text of each newspaper article. Unreported results for the analysis of the titles of each article have not produced sufficient results, which is different from Hausler *et al.* (2018). Our initial assumption that the titles and the book reviews share a similar structure was not confirmed. It seems that the classifiers rather rely on the word structure of the whole text and assign the classes based on the word frequency, therefore more words generate a more stable output. At the same time, the results of the correlation analysis and the probit model results are to some extend unexpectedly. Our initial thought, that a focused test corpus should generate a better sentiment indicator was not confirmed. We draw this observation back to the fact that the smaller corpora rely on a much smaller number of articles.

While more robust testing is required to establish unequivocal superiority of these sentiments measures, one of the most important contributions of the paper is the applicability of the shown methods in niche areas such as real estate market intelligence. Analyses of real estate issues tend to depend heavily on hard economic and property

market data or interviews and surveys. Understanding what an economic agent is planning to act on is quite complex and tricky to establish through observed data or questions asked in interviews or surveys. Sentiment measures derived from observed data may not reflect the true attitude of the economic agents. While the measures that we have derived do not fully close the gaps, those can add additional explanatory power to any analysis of economic relationships. This is especially useful in a market like real estate where imperfections are common, biases are rampant and drawing inferences are clouded with concurrent and competing trends.

At the same time, the field is evolving quite rapidly, and new methods are introduced regularly. Even so, our paper proves that there is learning potential from the book reviews. A better, more distinct training corpus could help improve the classification of the algorithms. One of the shortcomings can be found in the unbalanced structure of the test and the training dataset, with regards to the depth of words used. It is not surprising, that the book reviews use a smaller universe. However, the algorithms incorporate this lag and are therefore unable to classify a majority of words used in the articles, which is an issue worth investigating in the future.

REFERENCES

Bai, X. (2011). Predicting consumer sentiments from online text. Decision Support Systems, 50 (4), pp. 732-742.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. The Journal of Economic Perspectives, 21(2), pp. 129-151.

Bram, J., & Ludvigson, S. C. (1997). Does consumer confidence forecast household expenditure? A sentiment index horse race. Federal Reserve Bank of New York Economic Policy Review 4, (2), pp. 59-78.

Breiman, L. (2001). RANDOM FORESTs. Machine learning, 45(1), pp. 5-32.

Case, K. E., Shiller, R. J., & Thompson, A. (2012). What have they been thinking? Homebuyer behaviour in hot and cold markets (No. w18400). National Bureau of Economic Research.

Case, K.E., Shiller, R.J., (1989). The efficiency of the market for single-family homes. The American Economic Review 79 (1), pp. 125–137.

Chen, C. C., & Tseng, Y. D. (2011). Quality evaluation of product reviews using an information quality framework. Decision Support Systems, 50 (4), pp. 755-768.

Chen, T., Xu, R., He, Y., Xia, Y., & Wang, X. (2016). Learning user and product distributed representations using a sequence model for sentiment analysis. IEEE Computational Intelligence Magazine, 11(3), pp. 34-44.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1), pp. 2-9.

Clayton, J., Ling, D. C., & Naranjo, A. (2009). Commercial real estate valuation: fundamentals versus investor sentiment. The Journal of Real Estate Finance and Economics, 38(1), pp. 5-37.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20 (3), pp. 273-297.

Das, P. K., Freybote, J., & Marcato, G. (2015). An investigation into sentiment-induced institutional trading behavior and asset pricing in the REIT market. The Journal of Real Estate Finance and Economics, 51(2), 160-189.

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: the case of REITs. The Journal of Real Estate Finance and Economics, 45(2), pp. 402-434

Fan, T. K., & Chang, C. H. (2011). Blogger-centric contextual advertising. Expert systems with applications, 38 (3), pp. 1777-1788.

Foote, C. L., Gerardi, K. S., & Willen, P. S. (2012). Why did so many people make so many ex-post bad decisions? The causes of the foreclosure crisis (No. w18082). National Bureau of Economic Research.

Freybote, J. (2016). Real estate sentiment as information for REIT bond pricing. Journal of Property Research, 33(1), pp. 18-36.

Freybote, J., & Seagraves, P. A. (2017) Heterogeneous investor sentiment and institutional real estate investments. Real Estate Economics., 45 (1), pp. 154-176.

Friedman, J. (1996). Another approach to polychotomous classification (Vol. 56). Technical report, Department of Statistics, Stanford University.

Hausler, J., Ruscheinsky, J., & Lang, M. (2018). News-based sentiment analysis in real estate: a machine learning approach. Journal of Property Research, 35(4), pp. 344-371.

He, Y. (2012). Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. ACM Transactions on Asian Language Information Processing (TALIP), 11(2), pp. 4.

He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. Information Processing & Management, 47 (4), pp. 606-616.

Heinig, S., & Nanda, A. (2018). Measuring sentiment in real estate–a comparison study. Journal of Property Investment & Finance, 36 (3), pp. 248-258.

Heinig, S., Nanda, A. & Tsolacos (2020). Which Sentiment Indicators Matter? Evidence from the European Commercial Real Estate Market. Journal of Real Estate Research.

Published online - https://www.tandfonline.com/doi/full/10.1080/08965803.2020.1845562

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks, 13(2), pp. 415-425.

Kumar, M. A., & Gopal, M. (2009). Least squares twin support vector machines for pattern classification. Expert Systems with Applications, 36(4), pp. 7535-7543.

Liaw, A., & Wiener, M. (2002). Classification and regression by random RANDOM FOREST. R News, 2(3), pp. 18-22.

Marcato, G., & Nanda, A. (2016). Information content and forecasting ability of sentiment indicators: case of real estate market. Journal of Real Estate Research, 38(2), pp. 165-203.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5 (4), pp. 1093-1113.

Min, H. J., & Park, J. C. (2012). Identifying helpful reviews based on customer's mentions about experiences. Expert Systems with Applications, 39 (15), pp. 11830-11838.

Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40 (2), pp. 621-633.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42 (24), pp. 9603-9611.

O'Keefe, T., Curran, J. R., Ashwell, P., & Koprinska, I. (2013). An annotated corpus of quoted opinions in news articles. In ACL (2) (pp. 516-520).

Preis, T., Reith, D., & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 368(1933), pp. 5707-5719.

Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. Decision support systems, 53 (4), pp. 754-760.

Sadique, S., In, F. H., & Veeraraghavan, M. (2008). The impact of spin and tone on stock returns and volatility: Evidence from firm-issued earnings announcements and the related press coverage. Available at SSRN 1121231.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. Information Processing & Management, 52(1), pp. 5-19.

Smales, L. A. (2016). News sentiment and bank credit risk. Journal of Empirical Finance, 38, pp. 37-61.

Soo, C. (2015). Quantifying animal spirits: news media and sentiment in the housing market. Ross School of Business Paper, (1200).

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), pp. 1139-1168.

Tsai, F. T., Lu, H. M., & Hung, M. W. (2016). The impact of news articles and corporate disclosure on credit risk valuation. Journal of Banking & Finance, 68, pp. 100-116.

Tsolacos, S. (2012). The role of sentiment indicators for real estate market forecasting. Journal of European Real Estate Research, 5(2), pp. 109-120.

Walker, C. B. (2014a) Media and Opinion Leaders in the Housing Market, Queen's University Belfast, Working Paper FIN 14-8

Walker, C. B. (2014b). Housing booms and media coverage. Applied Economics, 46(32), pp. 3954-3967.

Walker, C. B. (2016). The direction of media influence: Real-estate news and the stock market. Journal of Behavioural and Experimental Finance, 10, pp. 20-31.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., & King, J. (2012). That is your evidence?: Classifying stance in online political debate. Decision Support Systems, 53 (4), pp. 719-729.

Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. (2011). Fine-grained sentiment analysis with structural features. In Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 336-344.

Table 1 - Descriptive statistics for the newspaper corpus on a quarterly level

	Full corpus	London	Newspapers +100k	without housing	FT
Mean	2,273	1,367	1,099	553	249
maximum	3,182	2,643	1,533	816	520
minimum	1,800	-	853	447	145
Sum	109,103	65,617	52,741	26,521	11,948
Range	1,382	2,643	680	369	375
standard dev.	375	775	156	81	102
Variance	140,492	600,959	24,354	6,564	10,502

Table 2 - Sparse Matrix (Newspaper articles)

	Amazon book reviews	Newspaper articles
Document Term Matrix	(documents 224,394, terms 580)	(documents 109,103, terms 5,354)
Non-/sparse entries	3,737,916/ 126,410,604	27,080,166 / 557,057,296
Sparsity	97.00%	95.00%
Maximal term length	13.00	19.00
Weighting	term frequency (tf)	term frequency (tf)

Note - The sparsity matrix illustrates the distribution of words in the used document corpus. It further summarises the total number of documents in the corpus and the total number of terms. Sparsity indicates how much of the matrix is empty (0).

Table 3 - Descriptive statistics for the dependent variable

Binary Capital Growth series	All Assets all properties	Offices in London Mid-Town and West End
Percentage of observations with negative growth	29.17%	17.36%
Obs.	144	144
Mean	0.292	0.174
Std. Dev.	0.456	0.380
Min	0	0
Max	1	1

Note - The table provides the descriptive statistics of the MSCI capital growth rates between 2004m1 and 2015m12.

Table 1 - Correlation matrix RICS vs. Textual sentiment indicators

	UK RICS PROPERTY SURVEY: SALES & RENTAL LEVELS-LONDON, NEXT QTR	*UK RICS SURVEY: OFFICE SALES & RENT LEVELS-LONDON, NEXT QTR NADJ*	*UK RICS SURVEY: RETAIL SALES & RENT LEVELS-LONDON, NEXT QTR NADJ*
SVM (all articles)	***0.575***	***0.586***	0.448
SVM (no housing)	0.175	0.122	0.212
SVM (London)	***0.558***	***0.542***	0.489
SVM (circulation above 100,000)	0.481	***0.502***	0.405
SVM (Financial Times)	-0.011	-0.001	-0.007
MAXENT (all articles)	***0.635***	***0.660***	0.487
MAXENT (no housing)	0.425	0.385	0.417
MAXENT (London)	0.296	0.215	0.374
MAXENT (circulation above 100,000)	0.479	0.481	0.432
MAXENT (Financial Times)	0.067	0.063	0.108
Random Forest (all articles)	***0.573***	***0.613***	0.419
Random Forest (no housing)	***0.603***	***0.604***	0.493
Random Forest (London)	0.481	0.405	***0.514***
Random Forest (circulation above 100,000)	0.514	***0.535***	0.443
Random Forest (Financial Times)	0.273	0.272	0.244

Table 2 - Probit model (MSCI converted capital growth all assets all properties)

Dependent variable: Change of the MSCI all assets all properties series

VARIABLES		(1) SVM (all articles)	(2) MAXENT (all articles)	(3) Random Forest (all articles)
z_svm_all = L,	standardized values for SVM all articles measure	-0.278**		
		[0.114]		
z_maxent_all	standardised values for MAXENT all articles measure		-0.317***	
			[0.115]	
z_rf_all = L,	standardised values for Random Forest all articles measure			-0.223**
				[0.100]
Constant		-0.558***	-0.564***	-0.556***
		[0.113]	[0.113]	[0.112]
Observations		143	144	143
Log likelihood		-83.43	-82.8	-84.22
LR Chi2		6.298	8.250	4.710
Lag		1	0	1
pseudo-r-squared		0.036	0.048	0.027
AIC		170.857	169.597	172.445
BIC		176.782	175.537	178.371
Correctly classified (%)		71.330	72.220	70.630
Sensitivity		4.760	9.520	2.380
Specificity		99.010	98.040	99.010
Hosmer-Lemeshow chi2		6.440	6.090	8.590
Prob > chi2		0.598	0.637	0.378
area under Receiver Operating Characteristic (ROC) curve		0.627	0.620	0.665

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)

Table 6 - Probit model (MSCI converted capital growth all assets offices in Mid Town and West End series)

Dependent variable: Change of the MSCI all assets offices in Mid Town and West End series

VARIABLES		(1) SVM (all articles)	(2) MAXENT (all articles)	(3) Random Forest (all articles)
z_svm_all = L,	standardized values for SVM all articles measure	-1.061***		
		[0.245]		
z_maxent_all	standardised values for MAXENT all articles measure		-1.080***	
			[0.204]	
z_rf_all = L,	standardised values for Random Forest all articles measure			-1.345***
				[0.258]
Constant		-1.146***	-1.166***	-1.215***
		[0.156]	[0.157]	[0.168]
Observations		144	144	144
Log-likelihood		-49.33	-44.26	-44.68
LR Chi2		34.26	44.42	43.58
Lag		0	0	0
pseudo-r-squared		0.258	0.334	0.328
AIC		102.668	92.51593	93.351
BIC		108.607	98.45555	99.29086
Correctly classified (%)		86.110	88.190	86.810
Sensitivity		28.000	44.000	44.000
Specificity		98.320	97.480	95.800
Hosmer-Lemeshow chi2		13.230	9.210	15.020
Prob > chi2		0.104	0.325	0.059
area under Receiver Operating Characteristic (ROC) curve		0.820	0.870	0.866

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)