Norwegian School of Economics
Bergen, Fall 2021

Sentiment Analysis in The Norwegian Housing Market

Evaluating the inferential and predictive power of sentiment scores on housing price using linear modelling and machine learning

Mads Parr Yksnøy, Erik Skutle

Lars Jonas Andersson

Master thesis, Economics and Business Administration

Major: Business Analytics, Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgements

This thesis has been written during the fall of 2021 at the Norwegian School of Economics (NHH), as part of our MSc degree in Economics and Business Administration.

The work on this thesis has been challenging, but also very rewarding. Since one of us is majoring in Business Analytics and the other one in Finance, we believe that this thesis offers broad perspective on a research question that excites us both. We hope the readers find that our thesis illuminates some dark spots on the map, and gives a small contribution to the research on relationship between information, sentiment, and the housing market pricing.

We would like to express our gratitude to those that have contributed and guided us prior to and throughout the research process. First, we would like to thank our supervisor Lars Jonas Andersson for clear guidance and valuable insights during these last six months. We also wish to thank Sven Are Nydal at the NHH-Studio for providing us with a powerful computer, enabling us to work with large amounts of data on complex algorithms. Next, we would like to thank Eiendomsverdi for sharing a vast amount of data on housing transactions in Oslo. Lastly, we would like to thank Henrik Wolstad and Didrik Dewan for providing insight into the concept of sentiment analysis.

Norwegian School of Economics

Bergen, December 2021

Mads Parr Yksnøy

Erik Skutle

Abstract

In this thesis, we investigate how information and sentiment provided through news media affect prices in the Norwegian housing market. Our analysis is based on news articles from selected Norwegian news outlets, transaction data from the housing market in Oslo and macroeconomic data. We derive sentiment values from the news articles using a recurrent neural network algorithm. We infer on the data using an OLS regression model and study the predictive ability of sentiment using XgBoost on models with and without sentiment data.

We observe that the variation in measured sentiment values explains almost half the variation in the housing price index for Oslo. This suggests that people respond to the information provided in the newspapers, and that the price development is not a random walk. Further, we observe that the sentiment coefficient is significant both in statistical and economic terms after we control for fundamentals, suggesting that people react to sentiment more excessively than what is justified by the fundamentals. The implication is that the housing market is not fully efficient. This is supported by data showing that an increase in sentiment values also widens the difference between asking price and final price. With the introduction of the XgBoost model, we decrease predictive error present in linear regression predictive benchmark by 14.96 percent. Our best sentiment model causes a decrease in prediction error of 2.52 percent relative to the reference model. This leads us to conclude that both fundamental information and sentiment is associated with price developments in the Norwegian housing market.

Keywords – Template, NHH, master thesis, LaTeX

Contents

List of Figures

List of Tables

1 Introduction

"*If you don't read the newspaper, you're uninformed. If you read the newspaper, you're mis-informed.*" ***Mark Twain***

We live in a time with an unprecedented amount of information available, both through conventional media outlets and through the internet and social media. At the same time, there are continuous discussions on whether traditional media outlets are losing their relevance, and if they can be trusted. One can always ask people what they think, but observing what they do can be more revealing. In particular if they respond economically to relevant information: do they put their money where their mouth is?

One way to do this is to examine the housing market in Norway. On a weekly basis, a large number of newspaper articles about the housing market is published, and in the same period a large number of property transactions take place. We want to use this rich data-set for examining whether people respond to the information the articles provide, and how they respond. In addition, we want to investigate whether people respond rationally to the information provided, or if they tend to overreact to positive or negative news. Or more formally: how sentiment is affecting behaviour and prices.

With this short backdrop, we present our problem statement for our thesis:

How does information and sentiment provided through news media affect prices in the Norwegian housing market?

Textual analysis is no new discipline, but modern day computational power has opened the door for complex analysis of big data, including unstructured data such as text. Studies in this field are precious few, and to our knowledge no study that combines the use of sentiment, housing and macro data in a Norwegian setting, exists up until this point.

In this thesis, we will analyze data from almost 100,000 housing transactions in Oslo the last five years, over 8,500 relevant newspaper articles, as well as macroeconomic data. We start out analyzing the sentiment of the newspaper articles using machine learning techniques, extrapolating a sentiment value on a per day-basis. We then combine the sentiment data with housing transaction data and macroeconomic data, before investigating the relationship between the variables in the data using linear regression and machine learning.

1.1 Structure of the thesis

In this thesis we will first provide a description and a discussion of the relevant literature and theory related to our topic, both economic theories and theories on methodology. Next, we present the different data that has been collected for use in the analysis, the different variables that are derived from the data, and how the final data-set is formed. In section 4 we describe our choice of method in detail, through the design of both the sentiment analysis classification algorithm and the preparation and execution of our housing price predictors. In section 5 will present the findings from our analysis. Finally, we discuss our results and draw some conclusions in chapter 6 and 7.

2 Background And Theory

This chapter covers relevant literature and theory for our analysis. We begin by reviewing existing literature on market efficiency and housing price prediction, before discussing important elements to include when doing price prediction with sentiment analysis. This is followed by a theoretical section in which we elaborate upon concepts in and surrounding statistical learning. Lastly, we present the theoretical framework behind the models used in this thesis.

2.1 Literature Review

2.1.1 Sentiment and the Housing Market

Sentiment is broadly defined as the psychology behind investor beliefs (Keynes, 1936). In the same book, Keynes later describes it as: *(. . .) activities [that] depend on spontaneous optimism rather than mathematical expectations (. . .).* He later labelled this type of behavior that could not be justified by fundamental facts or mathematics as *animal spirit.*

The most common understanding of the term today – in an economic context – is any action or inaction that is explained by other factors than rational and calculated ones (Akerlof Shiller, 2009), such as emotions, gut feeling, believes, drives etc. Or in other words: any behavior that is not justified by the facts at hand.

The most quoted articles on sentiment analysis and the housing market are all using media articles as a basis for determining the societal levels of sentiment (Walker, 2014). In short, they utilize textual analysis to assess whether news articles are positive, negative, or neutral in relation to a certain topic (Feldman, 2013). The purpose of the analysis is to quantify the tone of voice of "mood" of text, so that one can compare and benchmark the sentiment for different purposes.

A problem with this approach is that newspaper articles are not just the bearers of sentiment, but also of fundamental information. If we observe a response based on particularly positive or negative news articles, one cannot immediately distinguish between the effects of the information and the possible effects of sentiment. The most common way to deal with this is to control for fundamental events and circumstances in the analysis,

thus leaving the sentiment coefficient to capture any effect that is not justified by the facts (Soo, 2015). This is an approach we will try out in our analysis later on.

2.1.2 Existing Studies on Inference and Prediction in the Housing Market

There are relatively few studies that directly examine the relationship between general news media sentiment and the housing market, but some have been carried out in the UK and in the US. We will discuss their findings and their applicability to the Norwegian housing market.

One extensive UK study looked into the relation between newspaper articles and housing prices, in the period from 1993 to 2008 (Walker, 2014). The study finds a significant relationship between sentiment measured from the news media, and the development of real house prices. His findings, supplemented by a study from 2012 (Brueckner, Calem, Nakamura, 2012), is concluded with a likelihood that the sentiment in the news media influences banks and lenders rather than home buyers directly. The article suggests that the relationship between media sentiment and house prices are caused by changes in credit supply, which again can cause shifts in the demand curve for housing.

It is reason to believe that the credit supply in Norway is less volatile than in the UK, since lending practices have been rather strictly regulated through special mortgage regulations the last decade (Regjeringen, 2021). These regulations ensure that objective traits with the consumer, such as income and debt ratio, are the key variables when assessing a mortgage application. Less so the general societal sentiment and expectations represented through the news media. Thus, the conclusion from Walker's paper might not be transferred directly to the Norwegian market.

Another paper further investigated the relationship between news articles and real estate (Walker, 2016). This time it is tested whether positive news about the real estate market and housing prices, affects the stock price of companies engaged in the housing market. The study finds that positive news is correlated with both the stock price and the trading volume of the stock. The findings illustrate that positive news about a particular industry affects the stock of companies in that industry, but does not say anything about the effects on housing prices.

An American study has also looked into the relationship between news media sentiment and housing prices (Soo, 2018). The study investigated the relationship in 34 cities in the United States, and found that newspaper articles had a significant predicative power for future house prices. However, the tendency was the news sentiment had much larger effect in areas where speculative investors were prevalent and where sub-prime loans were approved and taken out.

This suggests that the news media sentiment first and foremost influence groups that are extraordinary sensitive or attentive to the expectations of the future prices. The transaction costs are significantly higher in Norway than in the United States due to the stamp duty, and sub-prime mortgages are not an option in Norway due to the regulations mentioned earlier. Therefore, we believe speculative behaviour might be more prevalent in the United States than in Norway, and it is unclear if the findings also apply to the Norwegian housing market.

Another study conducted in the United States found that Soo's results were also valid for commercial real estate, as sentiment reflected in the Wall Street Journal predicted the price development up to four quarters in advance (Beracha, Lang, Hausler, 2019). As buyers of commercial real estate usually are more interested in financial returns rather than other considerations more relevant to regular house buyers, these results are not directly transferrable to the Norwegian housing market. It could possibly be transferrable to the market for commercial real estate, but this is not a topic for our thesis.

Norges Bank – the Norwegian central bank – published a so-called staff memo in June 2021, examining how news media sentiment predicts housing prices (Kirkeby Larsen, 2021). They find a significant positive relationship in some of their applied models, but their focus is to examine how this dynamic works during economic turnarounds and crises. They use the covid19-pandemic as their backdrop, only predicting prices for December 2019 to March 2021. The memo is not peer reviewed, only uses news articles from one newspaper – Dagens Næringsliv – and does not distinguish between substantial front-page news, and minor articles. These factors make it difficult to transfer their findings to the general Norwegian housing market as such.

To summarize the existing literature on the topic, there are some studies that finds a positive relationship between news media sentiment and the development of real estate

prices. However, due to the circumstances we have pointed out, the results cannot necessarily be transferred directly to Norwegian housing market

2.1.3 Market Efficiency and The Housing Market

Generally, a market is efficient if prices fully reflect the information available, implying that new relevant information must lead to a price change (Fama, 1970). In a famous paper, the Efficient Market Hypothesis (EMH) is divided into three forms, weak, semi-strong, and strong. In the weak form of the EMH, the only relevant information that affects the price is the historical price. If the weak form of the EMH is true, the price development over time would simply be a random walk. If the semi-strong form om the EMH is true, the price would reflect the publicly available information, such as quarterly reports and similar. If the strong form of the EMH is true, the prices would in addition reflect monopolistic information that are only available to certain groups or individuals, or in other words: all information, both public and private.

The pricing of securities follows a simple logic; the value of a security today is the present value of the expected future cashflows from the security. The housing market works differently. You still have financial investors who values properties the same way the value securities, but the main group of buyers and sellers are regular people. One can assume that they also consider the financial implications of buying a home, but other considerations such as suitability for the family's needs are also present. This makes it a little bit more complicated to empirically test the EMH in housing market, but the basic idea remains the same: If prices move without fundamental information – financial or non-financial - justifying it, the housing market is not efficient – and vice versa. If the housing market is efficient, the price equation could formally be written like this:

$$P(House) = PV(FutureCashFlows) + Non - FinancialUtility + \epsilon$$

Several scholars have done empirical research on the efficiency of the housing market, and the most cited ones (Capozza Seguin, 1994; Pollakowski Ray, 1997) conclude that the prices do in fact change somewhat as new information becomes available, but not so that that the change is fully consistent with market efficiency. They explain these results partly by pointing to substantial transaction costs. This makes sense, since in addition to

the financial costs of a real estate agent and possible stamp duties, physically moving to a new home comes with substantial non-financial costs such as time and effort, making the threshold higher for acting on information.

2.2 Theory

2.2.1 Textual- and Sentiment Analysis Fundamentals

Sentiment analysis is the study of analyzing opinion and sentiment towards entities, such as products, services, etc., through text (Agarwal et.al, 2016). Two types of methods have been used in the literature for sentiment analysis. The first one is the machine learning approach while the second is the so-called semantic orientation approach. Sentiment analysis classification using machine learning usually face some challenges. One of them being that machine learning approaches produce high-dimensional feature vectors consisting of noisy, irrelevant and redundant features. Most of the existing feature selection techniques, used for sentiment analysis, do not consider the redundancy among the features. Existing methods select the important features based on goodness criteria for the class attribute. Traditionally, another problem has been that generated feature vectors have had problems with sparsity of data.

The latter approaches using semantic orientation are categorized into corpus-based and lexicon-based (knowledge-based) approaches. According to Agarwal (2016), Corpus-based approaches mainly depend on the method to determine the polarity of the words used. These approaches do not perform well because polarity of words changes with the domain and context, and there is no text corpus available which can provide polarity of words depending on the domain and context. Knowledge-based approaches depend on the already developed knowledge bases like SentiWordNet, WordNet, etc. The problem with these approaches is the type of coverage, as most of the available knowledge bases contain general knowledge (not contextual knowledge). General knowledge is often insufficient when determining the polarity of the document. The sentiment values employed in our analysis are calculated through the use of machine learning techniques.

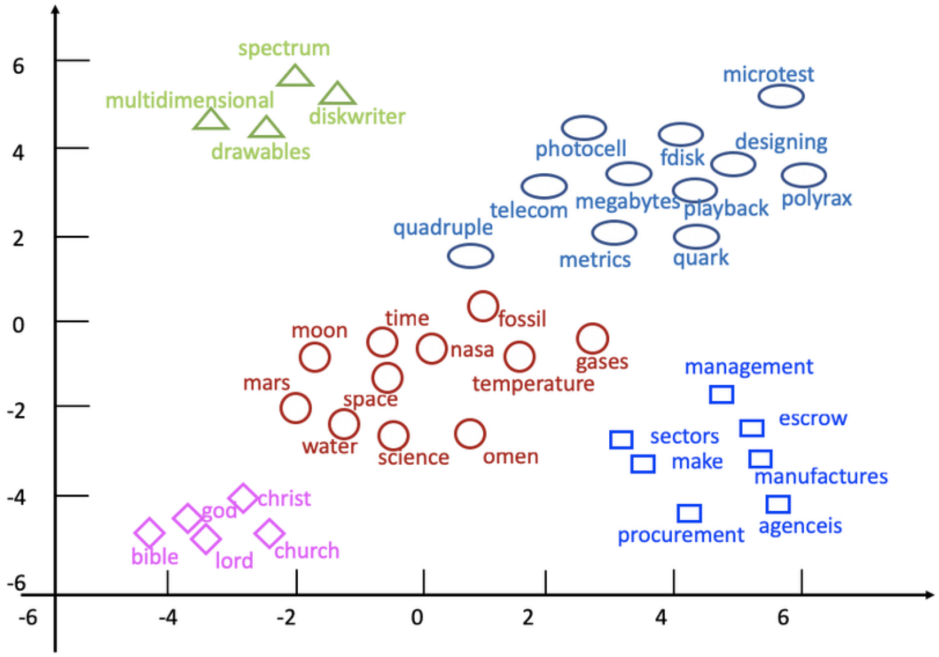

Figure 2.1: Word Embedding Illustration (TMSA, 2019)

2.2.2 Word Embedding

Within the field of sentiment analysis, word embedding is a term used for the representation of words for text analysis (Almeida, 2019) . It typically takes the form of a vector, where words are encoded into numerical representations, and where words that are similarly used in the data are placed closer together in the space. Figure 2.1 illustrates how words used in the same context are placed closer together in space. In the context of data handling, mapping of words is done through the use of a matrix structure.

Word embedding has some limitations in that the method of converting words into number results in the merging of a word that might otherwise have several meanings based on context, into one numerical representation. This causes the word to lose some of its contextual meaning.

2.2.3 Modelling for Inference and Prediction

When observing a quantitative response Y with p different predictors X1, X2,..., Xp. we assume that there exists an inherent relationship between the response variable and its

predictors (James et.al, 2013). This assumption can be formulated as

$$Y = f(X) + \epsilon \tag{2.1}$$

Statistical learning refers to a set of approaches for estimating *f*. Here, f is some fixed but unknown function of x1,...,xp,. ϵ is a random error term and signifies all the noise and all the movement that is not captured by *f*. The estimation of f is done for two main reasons: *inference* and *prediction*.

Inference focuses on how Y reacts to changes in its explanatory variables. The emphasis is here on understanding the relationship between the Y and X variable. As an example: How much of housing price can be explained by one extra square meter added to an apartment? Understanding the relationship between the two variables also includes understanding where the modelled relationship does well and where it falls short. To understand this relationship we need an open and interpretable model design, and in practice this excludes machine learning methods from inference use. Machine learning has been said to function like a black box, giving little insight into the form of *f*.

Prediction focuses on producing an estimation Y given a set of inputs X (James et.al, 2013). It is best applied in scenarios where there is a scarcity of output Y . As an example: in a pandemic scenario, how well does factors such as gender, age and underlying health conditions do at explaining the rate of death? Are there features that can be added in order to improve the prediction? While causality is something that is often discussed within the scope of inference, it is not something that is prioritized within prediction. The main purpose of predictive methods is to identify and gather a set of predictors that produce the most accurate predictions for output Y. The nature of the relationship and the form of $\hat{f}$ is less important. Machine learning techniques typically thrive on predictive methods.

2.2.4 Machine Learning

Machine learning is the process of using statistical tools to learn from data (James et al., 2013). These statistical tools are divided into two main categories: supervised and unsupervised learning. Supervised learning is the process of relating a response variable to a set of predictors. In other words, we use x to predict y to get a better understanding

of the relationship between the two.

Unsupervised learning does not focus on the output variable y. The process is based around understanding and exploring the relationship between the variables. In this thesis an output y will be employed, as the inference and prediction of housing prices falls within the category of supervised learning.

2.2.5 Data Partitioning

Prediction through the use of machine learning is based around developing a model that makes accurate predictions on new and unseen data (James et.al 2013). If the model is not applicable in a scenario where it is presented with new data, it has no real practical value. Facilitating of generalizability starts with data partitioning.

Data partitioning is the process of separating data into two or three data sets. One training set, one test set, and optionally, one validation set. The training set provides the machine learning algorithm with combinations of response- and explanatory variables, enabling training - and learning of the relationship of the data sources. The validation set is optional and is used as a way to provide quick feedback on model performance, and is as a result often used as a tool for hyperparameter tuning (Sarkar, 2016). It is best used in settings where data is plentiful. The test set is used to evaluate the performance of the final model, on independent- or out-of-sample-data.

2.2.6 Overfitting and Underfitting

The phenomenon where models follow the errors or noise to closely is called overfitting (James et.al, 2013). This usually happens when a model is fitted too close to the training data of the machine learning process, so that the model lacks the flexibility to perform well on new data with that do not follow the patterns of the data on which the model was trained on. This happens because the learning procedure works too hard to find patterns in the training data, thus picking up patterns that are simply caused by random chance rather than patterns that are caused by true characteristics of the variables in the model.

A simple analogy would be “teaching for the test” in schools, where teachers focus only on the type of questions they know will show up on an upcoming test. This may cause

the pupils to perform well on the test, but perform poorly in situations related to the underlying topic later on.

A typical sign of overfitting is when a prediction that has been fitted[1] to both a training set and a test set returns a lower prediction error on the training set than the test set. A model can also underfit. Underfitting is when the model fails to capture the relationship between the predicting and explanatory variables. A model is optimally fit when it is neither underfitting nor overfitting, as shown in figure 2.2.

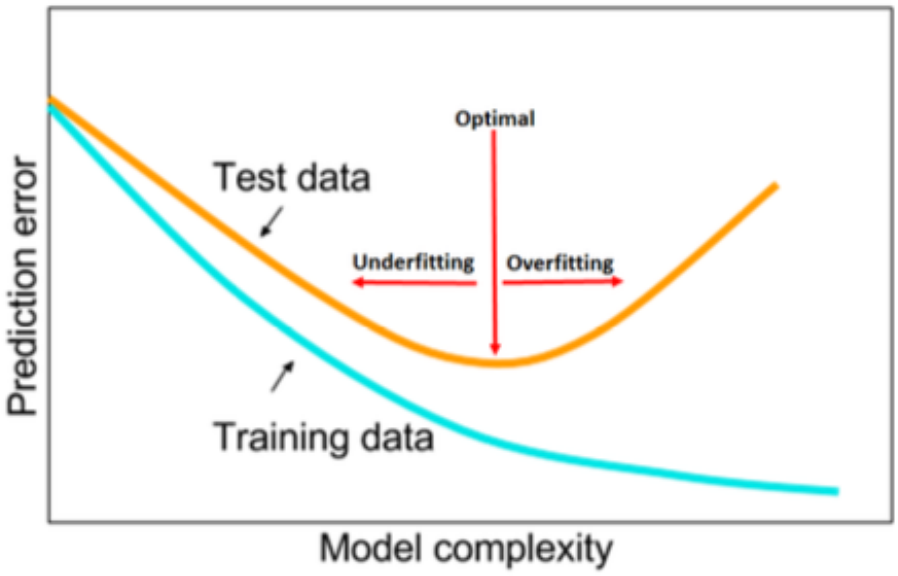

Figure 2.2: Optimal capacity between underfitting and overfitting(Kumar, 2021)

2.2.7 Bias-Variance Trade-Off

When seeking to minimize the model test error, we wish to utilize a statistical method used in the estimation of a model needs to be simultaneously low in variance and low in bias. In statistics, bias is the systematic error that a learning algorithm is expected to make when trained on training sets (James et.al, 2013). Learning algorithms are made to learn from the patterns that emerge in the training data and are as a result designed to adopt bias in order to generalize beyond the training data. Bias is also differently defined as a model´s shortcomings when modelling a complex data relationship. As an example, a linear regression model will introduce a large amount of bias onto a model where non-linear relationships exist. It is not flexible enough to capture the intricacies of the data. Variance is different. Variance refers to the amount by which $\hat{f}$ would change if

it was estimated on a different set of observations.

As a general rule, more flexible methods yield higher variance and lower bias, and vice versa. As one increases the flexibility initially, the bias tends to decrease faster than the variance increases, resulting in a lower prediction error. At some point, an increase in flexibility has limited impact on the bias, but strong impact on the variance, resulting in higher prediction error. The relationship between prediction error, variance and bias is referred to as the bias-variance trade-off.

2.3 Models

2.3.1 Linear Regression

The first model we use is an Ordinary Least Squares (OLS) regression. In it simplest form, such a model involves predicting a response variable Y, based on an explanatory variable X. It can however easily be extended to incorporate more explanatory variables for the same response variable Y. The fundamental idea of OLS and linear regression in general, is to fit a line through the data-points in such a way that the distance between the line and the data points is minimized (Woolridge, 2014). If the regression is a multiple linear regression the object fitted to the data points will not be a line, but a hyperplane. The advantage of OLS it that you get coefficients - estimations of X - that are easy to interpret, and thus easy to build a discussion on. The main disadvantage is that it simplifies the reality by assuming that all relationships are linear or quadratic. This might be true on average, but seldom for an individual line of data.

2.3.2 Neural Networks

2.3.3 Feedforward Neural Networks

In order to give some insight into the general nature of neural networks, we use the feedforward neural network as a tool to study networks, before studying the recurrent neural network. Neural networks originated in the computer science field and were designed to answer questions traditional approaches used in statistics were not optimized to handle. Inspired by the human brain and its capability for pattern recognition, interconnecting neurons processing information has been the ground for development of the artificial

neural network (ANN) (Wang, 2003). Among the artificial neural networks, we find the feedforward neural network (FNN).

The structure of the neural network is on the surface a simple one. One input layer of neurons, one to several hidden layers also made up of neurons, and a final layer of output neurons (Gupta, 2017). The layers are placed in a grid-like structure and connections are made between each node in every layer to every node in the layer succeeding the current one. Data is fed into the input layer, and is transformed in the hidden layers. For standard regressions, the number of output layers is locked at one, and this layer outputs a predicted value. In figure 2.3, we show a classic ANN with one hidden layer.

Figure 2.3: DFNN with one hidden layer(UC Business, 2019)

The input layer is in figure 2.3 represented by the four green circles. The number of units in the input layer is determined by the number of unique explanatory variables in the data set. The majority of all neural network learning takes place in the hidden layers, and by adding more than one hidden layer, more complex data interactions can be learned. The neural network consisting of more than one hidden layer is what determines whether the neural network is defined as deep or not. In the figure, the hidden layer is represented by the five purple circles.

Wang 2003 illustrates a typical neural network architecture through the mathematical representation in figure 2.4. The output h_i, of neuron i in the hidden layer is,

$$h_i = \sigma\Big(\sum_{j=1}^{N} V_{ij}x_j + T_i^{hid}\Big),$$

Figure 2.4: Neural Network Formula

Here, σ is the activation function, N is the number of input neurons, Vij is the weights, xj inputs to the input neurons and T_i^{hid} is the threshold terms of the hidden neurons.

Let us illustrate the concept of activation functions by using one type of activation function - Rectified Linear Unit function (ReLU) - as an example. This type of activation is best explained through the analogy of solar panels. A person living isolated from the outside

world obtains all his electricity from solar panels placed on the roof of his house. He can only run his washing machine on really sunny days. The activation function is dependent on "enough sun" or, put differently, a determinant of whether or not enough input has been given to fire a signal to the next layer of the model. This input is also determined by the element of weight, where each node-connection is assigned a weighting . Not included in formula x.x, a neural network normally includes an additional bias unit.

If the threshold has been reached, the activation function is employed. Through the keras package in R, we have access to a wide range of activation functions that differ in form and type, meant to capture different types of relationships. The ReLU function is a type of activation function that is most commonly used when attempting to capture the relationships in rectangular data. The function is simple, and binary. If the sum of the weighted inputs has reached the threshold, the ReLU-function returns a 1. If not, 0. This is formulated in figure 2.5

$$\texttt{Rectified linear unit (ReLU):}\ f(x) = \begin{cases} 0, & \text{for } x < 0. \\ x, & \text{for } x \geq 0. \end{cases}$$

Figure 2.5: ReLU Activation

The learning process of a FNN consists of adjusting the values of the weights between all the nodes so that the model fits the data well. Adjusting the weights is done in a manner that minimizes the loss function defined.

The exact method in which the weight values are adjusted is done through backpropagation in combination with a defined optimizing algorithm. When doing the initial run of the model, the DFNN will select some observations (a batch), and randomly assign weights across all the node connections before trying to predict the output. Backpropagation is the internal feedback signal that assesses the model´s own accuracy and adjusts the weights across the connections in order to try to improve accuracy. It is through the repetition of this process that the network learns the relationship between the data variables.

This thesis will implement the RMSProp(with momentum) optimizing algorithms. RMSProp is an extension of SGD which divides the computed gradient with a running average of its recent magnitude and employs an adaptive learning rate in order to allow it

to converge faster to an optimal solution.

2.3.4 Recurrent Neural Networks

The feedforward network architecture has been given its name on the basis of how the network processes information. Input flows directly through the hidden layers and subsequently becomes output. The recurrent neural network (RNN) works differently (West, 2019).

By adding a loop to the network hidden layers, the RNN is optimized for sequential processing. In practice this means that data is treated together by closeness in index or by date. This enables the model to account for development over time, or context. A line of text which is representative of how a large amount of sentences can be structured is:

"To say that the Norwegian housing market is cooling off would be a gross overstatement."

In order to make sense of the sentence above, each word must be interpreted with the words proceeding it in mind. By maintaining an internal state between separate inputs, the RNN is an excellent tool for text processing and classification. The RNN can be viewed as many copies of a Feed Forward ANN in a chain, as illustrated in figure 2.6

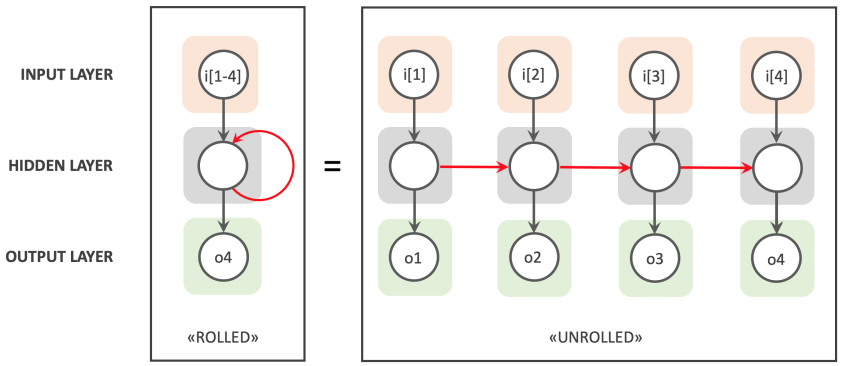

Figure 2.6: Recurrent Neural Network(West, 2019)

As mentioned in the chapter on FNN, the RNN also makes use of a backpropagation technique, only modified (Brownlee, 2017) . The backpropagation through time (BPTT) technique modifies the neural network weight over a sequence of timesteps$_1$. The network is unrolled, meaning that each affected input sequence is segmented into several parts, where timestep errors is calculated and accumulated. The network is then re-compressed and the weights are updated, before the sequence repeats.

Standard RNNs have been troubled with a limited "working memory", where input that appears early in a sequence and is important to overall context is "forgotten". This has been solved with the Long Short Term Memory (LSTM) network (Sinha, 2018). The LSTM network adds some extra components to the RNN hidden layer node. The new included components are the cell state, a forget gate, a input gate and a output gate. Figure 2.7 illustrates this.

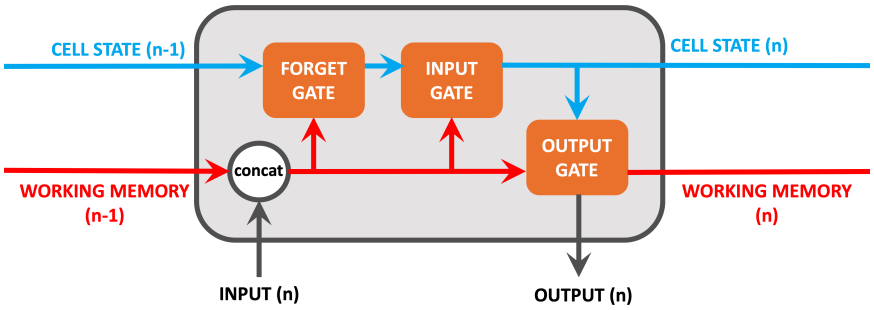

Figure 2.7: LSTM Hidden Layer Node (West, 2019)

The cell state works parallel to the working memory and is a second flow containing input over all iterations of the node, effectively maintaining what could become important contextual information. The forget- and the input gate works like a lock system, and links the two flows. If information from the cell state is deemed irrelevant to the long term flow, it is transported through the forget gate, removing its long term relevance. The function of the input gate is exactly the opposite. Information that could become relevant in the long term is transported through the input gate. The output gate calculates on which working memory the node will output.

LSTM RNNs have shown themselves to be better performing at sentiment analysis classification tasks than Deep Neural Networks and traditional RNNs. However, neural networks are largely dependent on proper tuning.

2.3.5 Neural Network Hyperparameter Tuning

Installing and running a neural network has in the last few decades become less computationally expensive, and a more manageable and available tool. While the implementation in and of itself has become easy enough, there are still a number of choices that must be taken in order to squeeze a network towards a maximized performance.

A neural network has many hyperparameter values that must be set. Some of these choices are determined on the data and the task the network must perform on the data, where the choice of regression or classification guides some of the hyperparameter choices. We will let established literature dictate some of the choices that are made, while the tuning process will be grounds for decisions on hyperparameters that are best optimized on a per-case-basis.

Table 2.1: Neural Network Hyperparameters

Hyperparameter	Description
Activation Function	The number of trees
Number of Layers	The number of trees
Number of Neurons	Maximum depth of a tree
Batch Size	Number of samples to draw from training data
Number of Epochs	Model learning rate
Learning Rate (LR)	Training set sample per tree
Dropout	Rate of weights to be dropped at each layer in each epoch
L2 Regularization	Complexity cost
Early Stopping Patience	Number of epochs with no loss improvement before training stops

How the neural network learns its non-linear features is contingent on which activation function the network utilizes (Glorot, 2010). The use-case of activation functions vary with classification and regression. In our classifying recurrent network, we will use the *softmax* activation function in the output layer. The softmax is optimized for multi-class classification cases such as ours. The ReLU activation function will be used in the RNN hidden layer and in the DFNN.

The number of layers determine the amount of hidden layers included in the model. Increasing the depth of a neural network is considered to improve the network ability to approximate functions with increased non-linearity. However, this comes with an increased risk of overfitting. Our RNN is made up of two layers.

The number of neurons in each of the hidden layers will be set by the tuning algorithm. If the number of neurons is too small, this may cause underfitting. If the number of neurons is too large it could cause overfitting.

The number of epochs sets the number of times the model iterates over the training set. The number of epochs that are actually iterated over is contained by the early stopping

parameter. By setting this parameter to 20, the number of epochs without loss is reduced, and the risk of overfitting through iteration is counteracted. Batch size is the number of training samples in each epoch. Small batch sizes gives a better model fit, while a larger batch size gives a better generalizing model.

The learning rate specifies how quickly a neural network learns. If the learning rate is too low, the learning process will be computationally demanding and converge slowly. If the learning rate is too high the model is likely to overfit. This value will be tuned.

A pattern that has likely emerged through the listing of different neural network hyperparameters, is their susceptibility to cause either overfitting or underfitting if not correctly set. Neural networks facilitate regularization through a handful of its parameters. Regularization is by essence a way of smoothening decision boundaries, improving generalization. Here, regularization will be employed through the use of the dropout and the L2 regularization parameter. Activating dropout will cause a number of outputs from hidden layers to be randomly ignored. By doing this, connectivity between nodes will be altered as a result of the shifting amount of nodes in the different layers. New connections will have to be made, and the network can learn more about the data. By activating L2 regularization, the penalty term "squared magnitude" to the loss function is introduced. Large inputs will be penalized and shrunk towards zero.

2.3.6 Extreme Gradient Boosting

The Extreme Gradient Boosting (XgBoost) is a tree-based model, where decision trees are used in model training and building (Morde, 2019). The decision tree method is an intuitive approach to supervised machine learning, and is applicable to both regression and classification problems. Each iteration of a decision tree model is often referred to as a *tree*. Figure 2.8 shows how a tree model is designed. Each tree is built from the ground up through a root node, where an initial "question" is asked. In a regression where the goal might be to determine housing price, the initial question asked could be "*How many bathrooms does the house contain?*". Depending on the amount of bathrooms in the house, the path (or branch) chosen will vary, just as if it were an *if-else* condition[1] . If the house has fewer than two bathrooms, choose the left branch. Else, choose the right one.

The question asking process continues throughout the depth of the tree model. When the tree runs out of depth and a leaf node is reached, a prediction will be given based on the conditions satisfied by the observation.

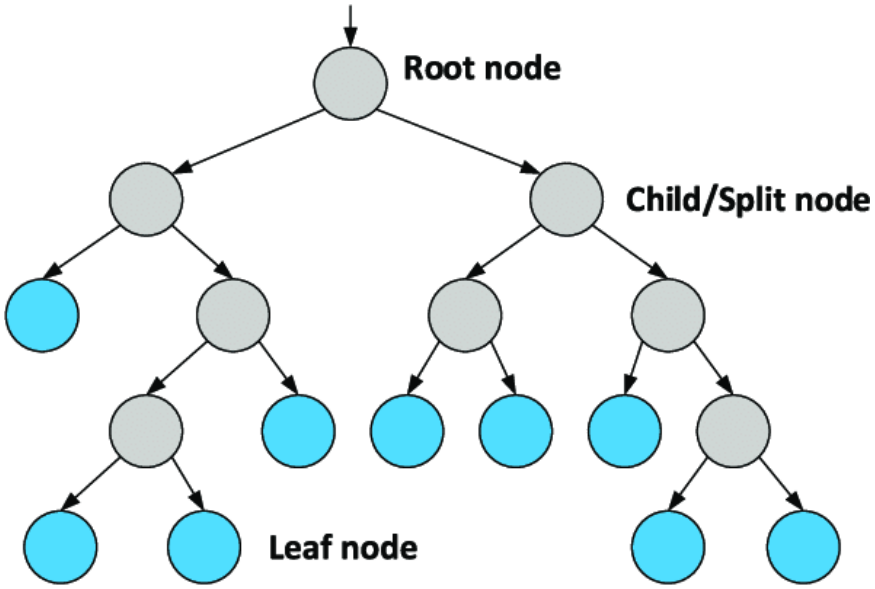

Figure 2.8: Decision tree layout with captions (Morde, 2019)

The extreme gradient boosting algorithm is based on the concept of the already established gradient boosting model. Gradient boosting refers to a type of ensemble machine learning algorithm. An ensemble is a collection of decision tree models, where trees are added one at a time. Information in the form of prediction error is then used to attempt to capture what makes a good predictor, and correct the model in succeeding trees. Learning from a set of iterations, turning a number of weak learners into one strong learner, is referred to as *boosting*. All the models are fitted through the use of are any arbitrary differentiable loss function and gradient descent optimization algorithm. The loss gradient is minimized during the model fitting process.

While built on the same principles, XGBoost introduces a more advanced and complete implementation of the Gradient Boosting algorithm. Among other additions, shrinkage and column subsampling are used to further prevent overfitting (Chen, 2016). A shrinkage technique is employ in order to combat overfitting and reduces the impact each fitted tree has in the model. It also makes room for the growing of new trees. Column subsampling is an alternative to the traditional row subsampling. Both methods increase variance between the tree models, and as a result allows the model to converge faster through

boosting, while preventing overfitting.

Shown in formula 2.2, the model tries to minimize the regularized objective using gradient boosting.

$$\zeta(\varnothing) = \Sigma l(\hat{y}_i, y_i) + \Sigma\Omega(f_k) \tag{2.2}$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

Here, ζ is a differentiable convex loss function. A convex loss function simplifies the process of reaching a global minimum, and makes it easier to find the best parameters globally. The formula measures the difference between $\hat{y}$ and target y. The second term Ω penalizes model complexity. The added regularization term helps to smooth the final learnt weight to avoid overfitting. The regularized objective will favor simplicity in its predictive models.

2.3.7 XgBoost Hyperparameter Tuning

XgBoost is also dependent on correctly set hyperparameters to optimize performance. Relative to the amount of hyperparameters that must be tuned for the neural networks, the tuning job of the xgboost model is simple.

Table 2.2: XgBoost Hyperparameters

Hyperparameter	Description
num-rounds	The number of trees
max-depth	Maximum depth of a tree
eta	Model learning rate
subsample	Training set sample per tree
gamma	Complexity cost
min-child-weight	Minimum sum of instance weight in child
colsample$_bytree$	sumsample ratio of columns

num-rounds is the number of decision trees used in the ensemble. Since each decision tree is added to the model in sequence and used to reduce prediction error made in previous iterations, increasing the number of trees usually improves model predictions. Increasing the number of trees can cause overfitting, and adding more trees will naturally increase computational time when training.

max-depth is an adjustment of tree depth and specialization. The choice between shallowness and depth gives a trade off between generalization and overfitting. Gradient boosting models typically perform best using modest depth.

eta is the model's learning rate. This parameter controls, intuitively, the rate at which the ensemble prediction learns from individual trees. A smaller learning rate will likely require an increase in the number of decision trees to gain the same amount of performance. However, by keeping a low learning rate, overfitting is reduced.

subsample is a measure of the number of samples that are used to fit each tree. By subsampling the data, we refer to a random selection of rows of data in the training set. A smaller sample set could cause a larger amount of variance in each tree, but improve the performance of the model as a whole.

colsample -bytree is similar to the the subsample parameter. Instead of sampling by row, the number of features (or columns) that are present in each tree is adjusted. This could increase variance in each tree, but give a better overall performance.

gamma or the "Lagrangian multiplier" controls the amount of model regularization. Gamma has been referred to as the complexity cost by introducing an additional leaf to the model. The larger the gamma value, the more a model is punished for its complexity. This is done to combat overfitting.

min child weight is somewhat similar to the gamma parameter. This parameter sets the minimum sum of instance weight needed in a child / split node. If a leaf node returns a instance weight sum that is lower than the value that has been set by the parameter, further partitioning of the tree is abandoned. This is done to reduce overfitting.

3 Data

In this section we describe all data that has been collected. We show how the data is finally combined, and present descriptive statistics of the data used in both phases of our analysis.

3.1 Data Sources

3.1.1 News Data

The first collection of data we cover is the news dataset. This data consists of news articles between January 1st of 2016 to December 31st of 2020. The data is obtained from the news database Retriever and contains just over 8500 news articles. All data were pre-sorted on-platform, by category as well as news provider. When settling on which providers to prioritize over this time period, we picked based on perceived ability to speak on Norwegian financial matters, and we focused on a total number of readers per newspaper.

We limited the search to the newspapers Verdens Gang (VG), Aftenposten, Dagbladet, Dagens Næringsliv and Finansavisen, as well as online articles from the online newspaper branches of NRK, TV2, and E24.

With these newspapers included in our search, we believe that the vast majority of the newspaper articles with potential to influence transaction decisions are included in our analysis. VG is Norway's largest newspaper with a circulation of 287,000 (UiB, 2021), Aftenposten is the second largest with a circulation of 257,000, and Dagbladet is the third largest with a circulation of 115,000. In addition, many of the articles are also published on their websites.

These are the three largest newspapers in Norway, and all have national coverage. However, the Oslo region does not have any local/regional newspaper of significant size. In our opinion, these three newspapers do in many ways serve as hybrids between national newspapers and regional newspapers for the Oslo region. The outcome seems to be that the real estate market in Oslo is covered very well in these newspapers.

Dagens Næringsliv is Norway's largest business newspaper with a circulation of 92,000, while their main competitor Finansavisen is more focused on financial news and has a circulation of around 24,000. NRK is the Norwegian public service broadcaster, with 1.1 million daily visits on their website, TV2 is the largest commercial TV channel in Norway with just above 1 million daily visits on their web page, while E24 is the largest online business newspaper in Norway with about 440,000 daily visits to their website (Hauger, 2019).

With their on-platform sorting mechanism, Retriever offers the option of selecting articles based on contained words. By constraining the data set to only display articles containing the words "bolig" or the word "eiendom", we efficiently decrease search results to only those relevant to our analysis. Then, the results are sorted further by news categorized as economics and business-news.

By default, exporting from Retriever gives us a txt-file with only one column. This contains all text found in each of the articles. To enable our analysis, we extract information on date and original publisher into separate columns.

If we analyze the distribution of articles over time, we see that the total volume of articles available in Retriever on the subject of housing has increased in the span of the available five years of data. In some of the earlier data, some days are not represented in the dataset, while many days are represented with as little as one data point. Early data also overrepresents some news-outlets to others, in contrast to the somewhat even distribution between sources we see in the later years. This could stem both from incomplete archiving from Retriever , or it could be a result of the subject of housing being more popular than before.

This data will ultimately be used as test data in our sentiment classification.

3.1.2 Review Data

We investigate the pre-labeled corpus approach for our sentiment analysis. This approach utilizes a large collection of naturally occuring text as the basis for analysis (Michigan ELT, 2010). Here, different pieces of text information have all been assigned a labeling, often on a scale of "bad" to "good". A bad rating and a good rating is mirrored by the words in the text belonging to the label.

Sentiment analysis training is often done with a lexicon-based approach, where documents containing a list of words labeled either as positive or negative contribute to addition or subtraction, respectively, from an overall sentiment score. However, this approach might oversimplify and lose some of the nuance found in the language. Machine learning-based approaches have been found to be more accurate in sentiment value predictions than their lexicon based counterparts (Nikil, 2019).

Knowing that our sentiment test data is written in Norwegian, we are dependent on comparable training data. For machine learning with the purpose of training and evaluating models for document level sentiment analysis, we gather data from The Norwegian Review Corpus (NoReC), provided by the Language Technology Group at Universitetet i Oslo (UiO). This data consists of full-text reviews from major Norwegian news sources and cover a range of different domains, including literature, movies, video games, restaurants, music and theater, in addition to product reviews across a range of categories. All reviews have been pre-labeled on the basis of a an already assigned review score. This means that each review has manually been assigned a score from 1-6, where 1 is poor and 6 is very good. We collect a total of 1200 reviews from this dataset.

3.1.3 Pre-Trained Word Embeddings

When we embed for natural language processing, the transformation creates a matrix where all the contained vectors represent words, and where words that are similarly used in text are placed closer together in the space. If we were to embed only our current training set of 1200 documents, our embedding would likely give a poor representation of where words should be placed in the space, due to the small sample size. Rare words would be more affected. To counteract this, pre-trained word embeddings are used. Pre-trained word embeddings are word embeddings that have been trained on large data sets in order to display the correct weight of each word in the matrix-space. For our analysis, we use three different pre-trained word embeddings from the NLPL word embeddings repository, created by the Language Technology Group at UiO. These are the "Norwegian-Bokmaal CoNLL17 corpus", "NBDigital" and "Norsk Aviskorpus + NoWaC"- corpus. All embeddings are without lemmatization, which means that the dataset retains the inflected forms of different words and do not reduce different inflected versions of a word into one word. All data is also based on a fasttext skipgram algorithm. What a skipgram is and

what it is not will be further discussed in chapter 4.1.1 and 4.1.4.

3.1.4 Housing Data

The *housing* dataset is received through an agreement with Eiendomsverdi, a real estate statistics and analysis company subsidiary to the association of real estate agents in Norway. The set contains data from almost 100,000 property transactions in Oslo the last five years. For each transaction the listing date, the sales date, asking price, final price, property size, postal code, age of the building, floor number of the property, size of the lot, property type, and ownership form is registered.

3.1.5 Housing Market Index

The housing market index is a publicly available price index produced by Eiendomsverdi. The index' starting point is January 2003, with a starting value of 100, and have since described the development of real estate prices in different Norwegian regions on a monthly basis.

3.1.6 Macroeconomic data

In addition to data from Eiendomsverdi and Retriever, we are also utilizing other publicly available data. Firstly, we are using data from Norges Bank on the key policy rate. We are using the nominal value of the key policy rate, as well as announcements on changes in the key policy rate in our analysis. The data is publicly available on the website of Norges Bank, but announcements have been transformed into binary data, based on press releases available on the website.

We also utilize published unemployment data from the Norwegian Labour and Welfare Administration (NAV). The NAV unemployment data is showing how many individuals that have registered as unemployed at NAV in the Oslo municipality.

Finally, we use monthly GDP statistics from SSB in some of our models. This data shows the percentage change in GDP per month, broken down based on region.

3.2 Developing the final dataset

In this section we will list the variables used in our analysis, and discuss the role of each variable.

3.2.1 Explanatory Variables

Sentiment The prediction output of the RNN-classification model is labeled "sentiment". It is an ordinal variable between -2 and 2, and is a measure of tagged sentiment, -2 being negative and 2 being positive.

We choose to introduce a linear decline to the sentiment variable. The choice to adjust the value of sentiment is based on research done on attention retention, which claims that the average reader spends only 15 seconds reading each article (TIME, 2014). We reduce the impact of sentiment values the further into the article we get in order to mirror the effect of a limited attention span.

$$S = \frac{\sum_{n=1}^{N}(N-n)\cdot X_n}{N} \tag{3.1}$$

Here, N is the total number of data rows in each article, while n is the index of the article row. Xn is sentiment value on line n. The values are summed and divided by the total number of rows per article, assigning an average sentiment value to the article as a whole. This new adjusted sentiment variable will take the place of the old one.

Since we want to capture the effect of sentiment on housing prices, we introduce differently lagged variations of sentiment, ranging from sentiment value on the day of the sale all the way to a 90 day lagged variation. All lagged values are rolling averages. The purpose of the lag is to investigate whether a possible effect is strongest in the shorter or longer run.

Table 3.1: Sentiment variable data

Variable Name	Description	Variable Type
Date	Time of transaction	Timestamp
SentimentDay	Sentiment, no lag	Score
SentimentL1	Sentiment, 1 day lagged	Score
SentimentL7	Sentiment, 7 days lagged	Score
SentimentL30	Sentiment, 30 days lagged	Score
SentimentL90	Sentiment, 90 days lagged	Score

To avoid duplication, the term "Justert" (English: "adjusted") is added to the variable name in the actual data-set, so that the variable "SentimentL1" is labelled "SentimentL1Justert" and so on.

3.2.2 Control Variables

While the main purpose of this thesis is to look at how sentiment affects price, we need to control for effects caused by other features. Our control variables will consist of both housing specific data and macro variables. Due to technical reasons, our variables have Norwegian name labels in the actual data set, but we will use English translations in the following discussion

BRA (Size)

Likely to be an import factor in how a property is priced, we use data on gross size of each of the properties present in the housing transaction data. This is measured in square meters.

Byggeår (BuildYear) Another important factor in determining the price is the year the property was built. It is hard to predict the effect this variable might have. Intuitively, one would assume that the newer a building is, the larger the positive effect on its price would be. However, if the city centre has been built from the centre and out - as Oslo has been-, the BuildYear variable might serve as a proxy for location and become a measure of how centrally the building is located. Seeing that our housing data does not contain any other location data, we have no other way of capturing the value of location. We think location within the city could be a strong predictor of prices.

D-renteoppgang (D-rateincrease) Since we want to isolate the effect of sentiment in and of itself, we would like to control for events that are likely to be discussed in news articles, but will have effects that go beyond psychological effects. *Dummy-rateincrease* is equal to 1 if there has been an announced increase in the key policy rate in the last 30 days, and is else set to 0.

D-rentenedgang (D-ratedecrease) For the same reasons as with the increase in the key policy rate, we wish to control for decrease. *Dummy-ratedecrease* is equal to 1 if a decrease in the key policy rate has been announced in the last 30 days, and is else set to 0.

Styringsrente (Rate) In addition to the effect of announcements of the key policy rate changing, we want to control for the level of the key policy rate. The level of the key policy rate affects the supply of capital directly, due to the Norwegian mortgage regulations, as well as affecting the demand for capital. An interest rate is the price of money, and if prices go up then demand is expected to fall.

Arbeidsledighet (Unemployment) As a control variable, unemployment statistics might serve multiple purposes. Firstly, it might capture some of the psychological effects that occur when a change in unemployment statistics is announced. Secondly, it can capture some of the effect of an income change in the population, and its effect on the aggregated demand for real estate. Thirdly, the unemployment level can serve as a proxy variable for various economic shocks or events that are difficult to control for directly, such as a pandemic.

BNP (GDP) For the very same reasons as with the unemployment statistics, we want to control for changes in GDP. We use the season adjusted change in GDP for mainland Norway from the previous month, measured in percent.

IndexKvarter (IndexQuarter) Finally, we wish to control for changes in price that occur as a function of time passing. We expect this variable to capture and control for effects of, among other things, inflation. Therefore, we group all observations into 60 different quarters based on the date the transaction took place.

3.2.3 Response variable

Pris (Price) Our analysis is based on inference and prediction of housing prices. This makes price a given response variable.

BPIndexO (HPIndexOslo) We also wish to explore how large amount of the movements in the housing price index can be explained by sentiment. We want to study the accuracy of sentiment where the response variable is aggregated prices by month and where property-specific features are removed.

Also,

Prisdiff (PriceDifference) *PriceDifference* is calculated from the difference between final price and asking price.

3.3 Final Data Subset

Collecting all of the variables gives the dataset seen in table 3.2.

Table 3.2: Variables in final dataset

Variable Name	Description
Date	Time of transaction
Response Variables	
Price	Housing price
HPIndexOslo	Monthly housing price index
PriceDifference	Difference between final price and asking price
Control Variables	
GRZ	Gross size per property
BuildYear	Year the property was built
D-rateincrease	Announcement of increased rate
D-ratedecrease	Announcement of decreased rate
Rate	Level of key policy rate
UnemploymentRate	Unemployment statistics
GDP	Change in GDP
IndexQuarter	Index for each quarter
Explanatory Variables	
Sentiment	Sentiment, no lag
SentimentL1	Sentiment, 1 day lagged
SentimentL7	Sentiment, 7 days lagged
SentimentL30	Sentiment, 30 days lagged
SentimentL90	Sentiment, 90 days lagged

3.4 Descriptive Data

We want to give a brief description of the final data-set, starting with a table that shows the number of observations, mean, standard deviation, and the maximum and minimum values of the variables used. Next, we display three histograms showing the frequency of the values in the price data, daily sentiment data, and the 90-day rolling average of the sentiment values.

The histogram for price looks a little skewed, but this is due to some extreme outliers: really expensive properties. The histogram on daily sentiment scores shows a spread of frequencies mostly between 0 and 5, and the histogram on the 90-day rolling average has a spread between 0 and 2.5. The main point of displaying these histograms is to show that we have a variation in our data which makes an analysis meaningful, as well as providing

some insight on how much "one unit of sentiment" is.

Variable	Obs	Mean	Std. Dev.	Min	Max
Pris	95,999	4989641	2858834	16985	7.11e+07
Prisant	95,933	4838963	2867126	250000	8.00e+07
BRA	95,369	77.32524	48.04952		1009
Prisdiff	95,999	154004.3	489327.8	-1.52e+07	4.40e+07
Byggeår	95,828	1957.57	42.81177		2021
Styringsre~e	95,999	.6143189	.4135891	0	1.5
Navledighet	95,999	13830.62	8513.274	9010	53453
BNP	94,729	.3516389	1.191661	-4.5	2.4
SentimentD~t	90,706	1.526069	1.108341	-.5502959	11.18227
Sen~1Justert	89,143	1.512186	1.052436	-.9847919	11.18227
Sen~7Justert	94,019	1.253035	.7077361	.0317937	3.320259
Se~30Justert	92,650	1.217854	.6289413	.2066402	2.789073
Se~90Justert	89,238	1.194559	.5877975	.34484	2.410889

Figure 3.1: Descriptive statistics

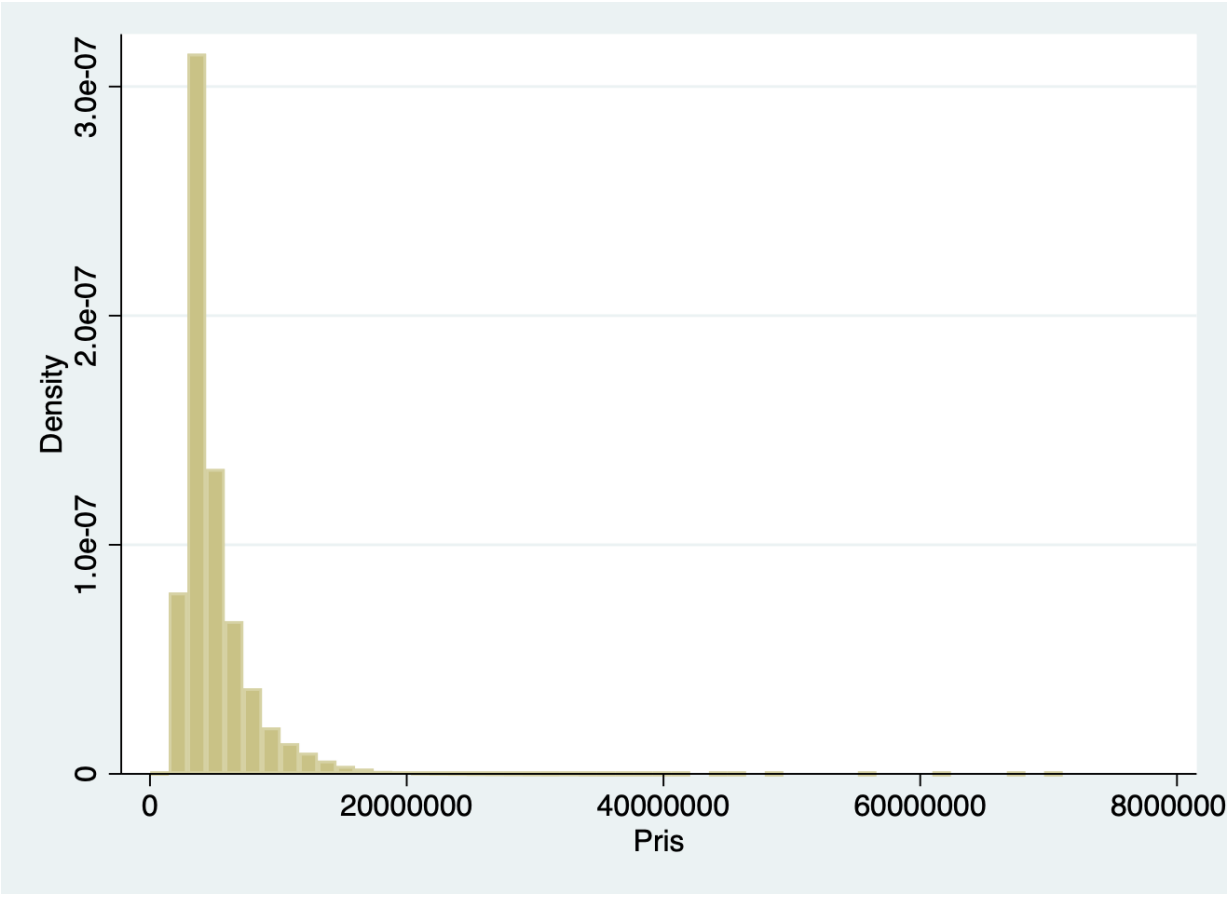

Figure 3.2: Histogram showing prices

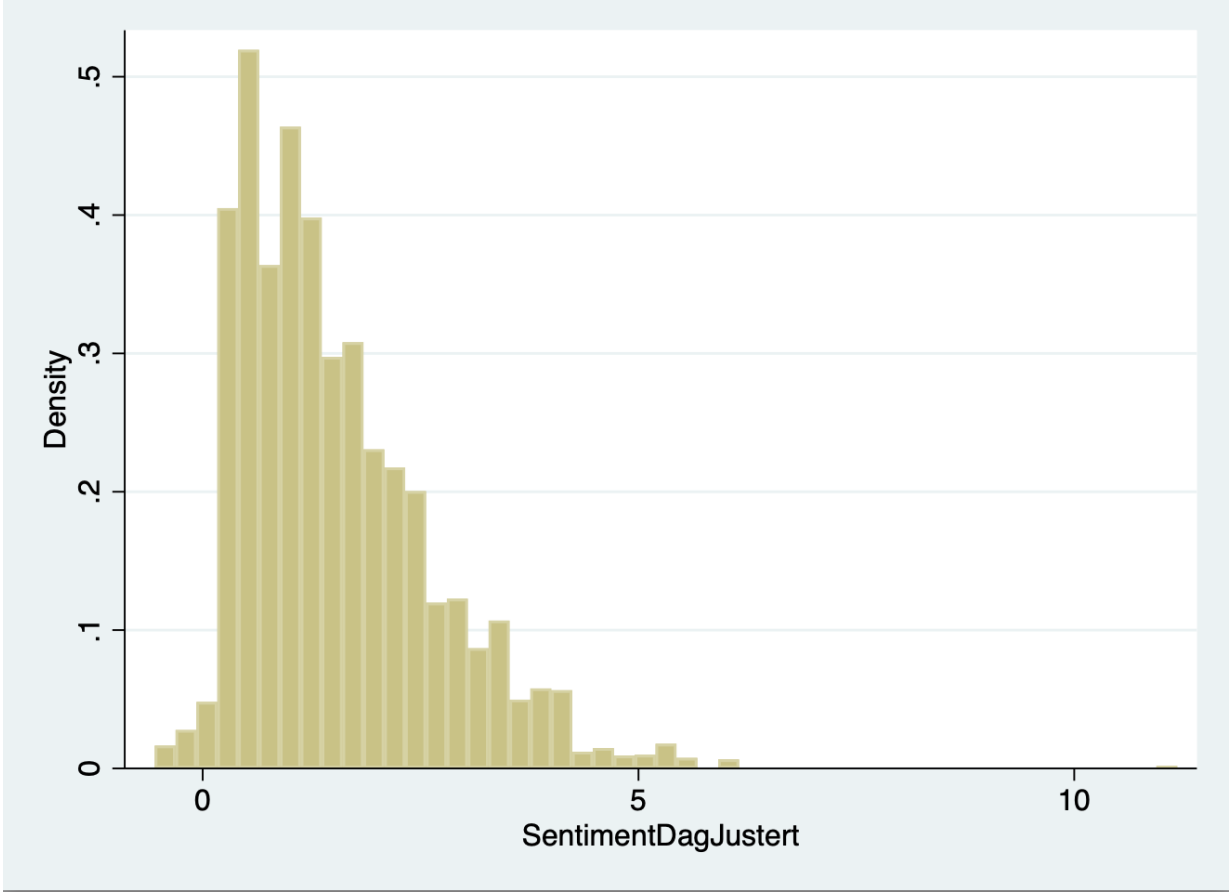

Figure 3.3: Histogram showing daily sentiment

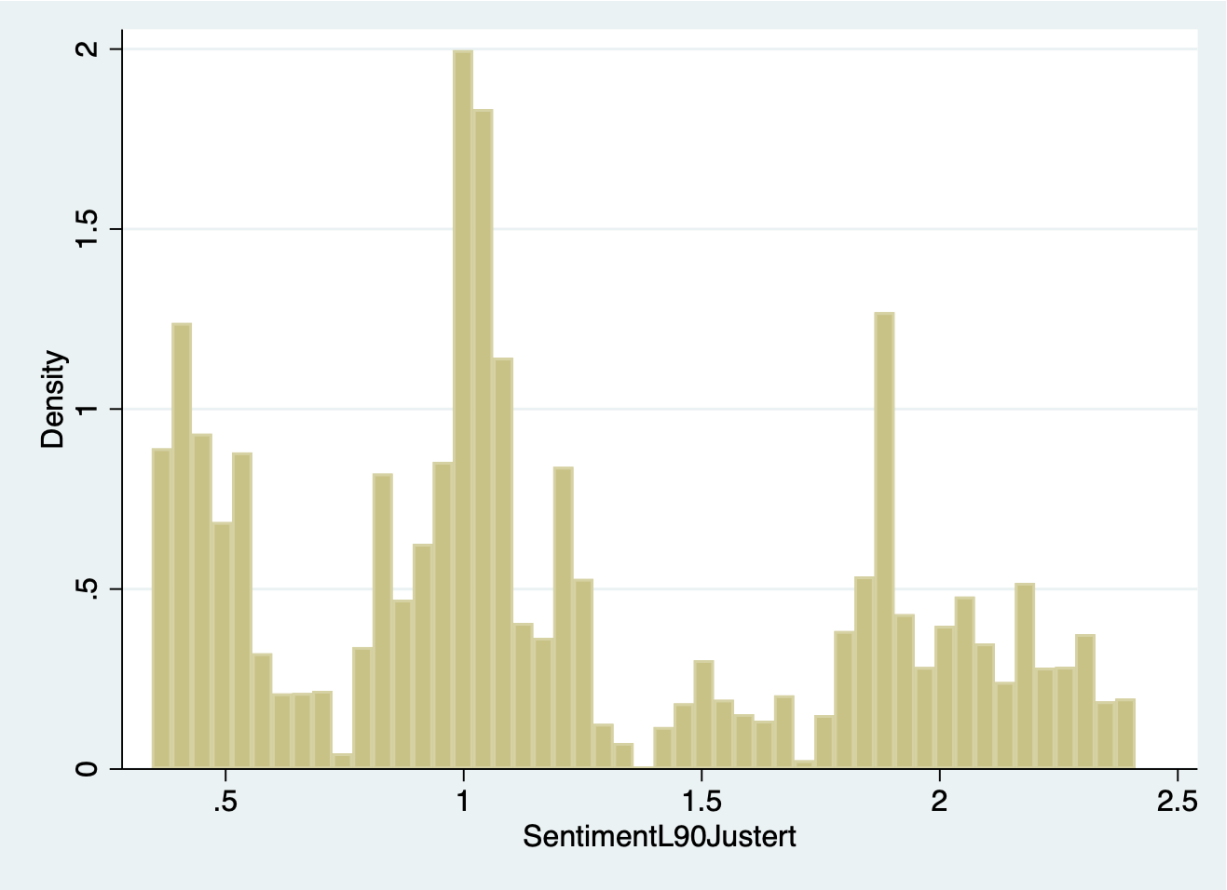

Figure 3.4: Histogram showing 90-day sentiment average

4 Methodology

This section is divided into three main part. First, the methodical approach to the sentiment classification algorithm is discussed. Then, the preparation and execution of housing inference and prediction is discussed. Lastly, machine learning hyperparameter tuning is discussed.

4.1 Textual Analysis

Textual analysis is a method using text in order to gain information. In the context of machine learning, the process starts with a source of text that can be analyzed and classified. Textual noise must be removed, and the data must be transformed so as to enable machine interpretation. The process ends when a model is able to assign a value to the initial source of text.

4.1.1 Pre-processing

Preparing and cleaning text for classification is in short called pre-processing (Haddi et al, 2013). What is required from the text cleaning process varies somewhat by how the text data is extracted. Gathering online text is often done in one of two ways. Either the text is extracted using a web scraping mechanism, giving the program direct access to the internet and enabling automation of data imports. This technique is often employed in programs where ease of use and quick response time are essential. An alternative to web scraping is a manual extraction of files. We chose to download all of our data manually, as this greatly reduces noise in the data and subsequent data cleaning.

Even though much of the noise is reduced through the use of manual extraction, the remaining data cleaning steps are important to data interpretability and in turn model accuracy and efficiency. Both sentiment- training and test data was processed by lowercasing all text, eliminating whitespace and by removing digits, punctuation and stopwords. We also excluded some of the additional pre-processing steps, like lemmatization. Lemmatization is a technique of reducing a word to its dictionary form while using word-context in order to determine its meaning. Applying lemmatization to our data resulted in several words being transformed into nonsensical stemmed versions

of themselves. This technique works well on English data sets, but the existing software is clearly not yet optimized to handle Norwegian text, which is why it is not used.

Each row, containing one article each, is then subsequently transformed into a skip-gram-structure. A skip-gram attempts to capture and predict the context for any given word in the data set. In practice, this changes the form of our initial "one article per row"- data table into a one - five words per row -data table, as seen in figure 3.1.

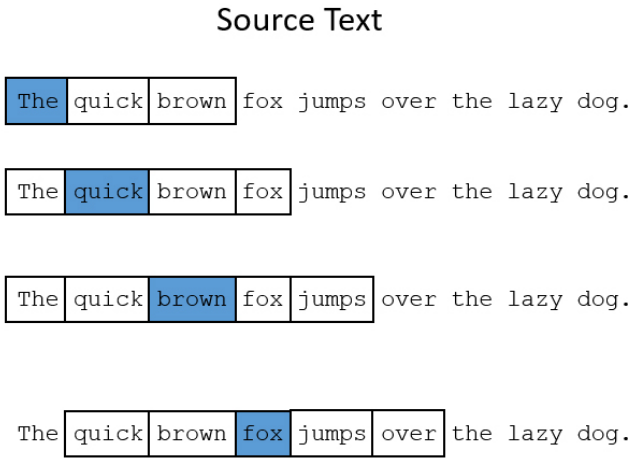

Figure 4.1: Skipgram Layout (McCormick, 2016)

4.1.2 Data Balancing

There are some considerations to be taken before feeding data into a machine learning model. Assessing the distribution of our classifiers is one of them.An imbalanced dataset is defined by an imbalance between the size of a minority class compared to the size of its majority class. Machine learning algorithms have traditionally been troubled by imbalanced data sets, since most expect balanced class distribution or an equal misclassification cost (Lemaitre, 2017). The absence of even distribution compromises the learning process, and makes correctly predicting minority classes much harder.

Using the pre-labeled corpus from NoReC, all training data is grouped into one of six groups of sentiment. Among the 4218 observations, groups that are classified as 1 and as 6 constitute less than one percent and six percent of the total. The data set is distributed as follows:

Table 4.1: Distribution of values before merge

Rating	1	2	3	4	5	6
N	42	234	716	1 460	1 510	256

The challenge of imbalanced data sets has many proposed solutions. One solution is to resample the data set, either through oversampling or undersampling, or both. Oversampling increases occurrences of the minority in the dataset, while undersampling decreases occurences of the majority class. One library, SMOTE, creates new synthetic data points of the minority class. This is a widely used technique, but performs poorly on text data, since the numerical vectors that are created are very high-dimensional and introduces data that overfits on the training data.

Undersampling our majority class is also not problem free, as most undersampling techniques are not developed for use on multi-class classification problems. We find that the best possible solution to the problem of uneven data distribution is to do a merge of the data labelled with a score of 1 with data labelled with a score of 2. As the distribution is now divided into an odd number of categories, this enables the middle category to be interpreted as a neutral category. Table 4.2 shows the new distribution. The distribution is still somewhat uneven, but better.

Table 4.2: Distribution of values after merge

Rating	2	3	4	5	6
N	276	716	1 460	1 510	256

4.1.3 Data Partitioning and Resampling

The validation set approach involves separating a randomly sampled segment of a training set into one new dataset: the validation set. While the ratio of the training and test sets of data is given a 80/20 split in favor of the training set, only 10 percent of the training set is separated into the validation set.

The model is then trained using the train dataset. After training, the fitted model will make predictions on unseen observations in the validation dataset. This way the model

can be evaluated based on out of sample data. The method is straightforward, easy to implement and will in some cases yield good results. However, the validation set approach is prone to overfitting, meaning the model is too closely fit to the train dataset and not performing well on new unseen data (James et al., 2013).

4.1.4 Word To Vector

An imperative part of natural language processing application is the embedding of textual information. Converting textual content into meaningful numerical representations enables machines to process and understand the text (Singh, 2019). We reduce the number of words to consider as features to 13000 unique words. Among the 14 900 unique words found in the dataset, these are the 13000 most frequently used, meaning that we remove the 2000 least used words from consideration. This is done in order to increase the relative importance of the remaining words, while also reducing computational time.

The data used in our analysis has been prepared through conversion to the skipgram-format. The skipgram is only one of many available conversion alternatives. One alternative is a continuous bag of words (CBOW). In contrast to the skipgram which seeks to predict the context based on the distribution surrounding, CBOW is used to predict the middle word in the skip gram; context predicting word. (Kulshrestha, 2019). The skip gram is slower trained, but outperforms CBOW in representation of infrequent words and phrases. However, a training set of 6000 articles is likely still not large enough to correctly represent the weight of most words in the corpus.

Also, with a limited sample size, generating self-made embeddings that generalise well to a test-set could prove challenging, and we are likely to encounter overfitting issues. To avoid this, we introduce our three pre-trained word embeddings. When training the model for the first time, we will freeze the weights of the model, hindering any updates from upper layers and a propagation up through into the training phase. The weights will be unfreezed for the second fitting of the model, allowing the training data to affect the weighting.

4.1.5 RNN Sentiment Classification

We start by encoding our initial input, using the three pre-trained word embedding weights. The resulting output of each of the embedding layers is then processed using a single long short term memory (LSTM) layer with a high dropout value. This is done to avoid an exponential growth by multiplying gradients through network layers with values larger than 1 (Nabi, 2019). Following this, the encoded embeddings are then merged together and condensed into a single layer.

The model is first trained on our validation data, then the model is used to predict on the test set. Each of the skip grams are marked by their corresponding documentID and matched by a sentiment value ranging from 1-5. Before aggregating the skip gram scores into one single document score, we re-adjust the sentiment value range by subtracting a value of three from each score. This shifts the sentiment values to a range of -2 - 2. By doing this, the effect of neutral statements on the prediction of data is eliminated. Another variable is then added to the dataset, combining the already established sentiment values with a linear value drop off over the course of one document. The new sentiment value is then used in computing a mean sentiment values for each news article. The data is then merged with the housing data provided by Eiendomsverdi.

4.1.6 Model Performance Metrics

In order to understand how sentiment values might drive house buy and sell decisions, the classification method must be a good reflection of how the housing market is discussed in public forums. A good reflection is contingent on a recurrent neural network that is able to accurately predict sentiment on test data, based on classifications made on the training data-set.

The training set's ability to accurately predict on the data-set will be assessed through the use of a validation set that makes up 20 percent of the total training set. Our model will be set to maximize classification accuracy on the validation set, where accuracy is measured between 0 and 1. An accuracy of 1 means the model is able to perfectly assign sentiment scores matching the sentiment scores already assigned the validation set. Validation

accuracy will also ultimately be the basis for the choice of RNN-hyperparameters, as the same metric is maximized in the hyperparameter tuning process. RNN classification performance can be seen in figure 4.2. The algorithm achieves a classification accuracy of almost 50 percent on the training data. In light of this being multi class classification, a high categorical accuracy prediction would be harder to achieve relative to binary classification.

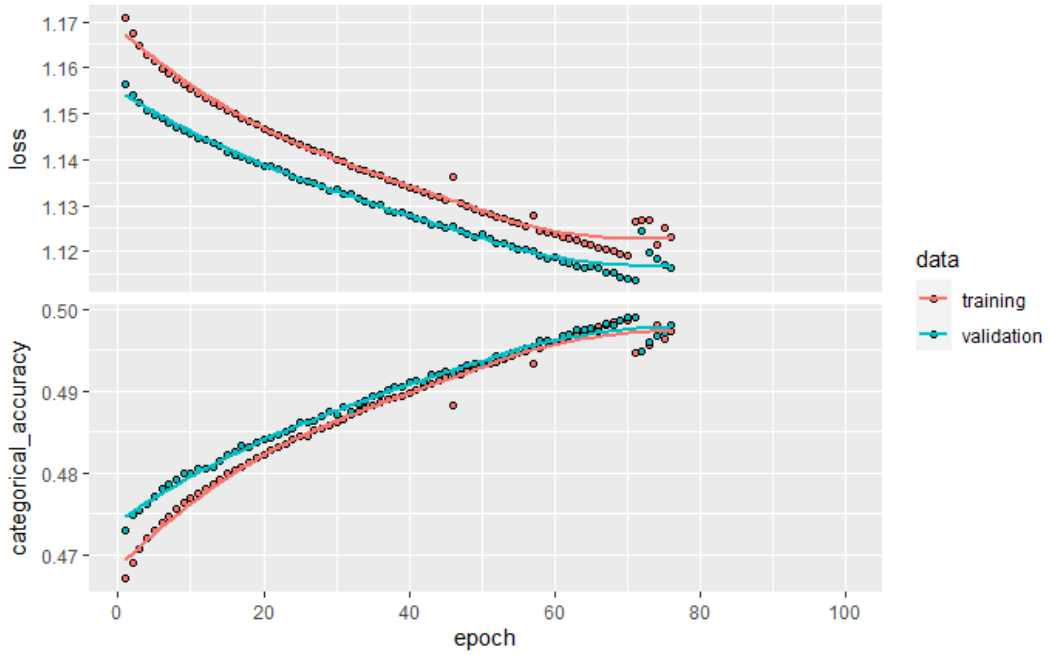

Figure 4.2: Performance of the RNN classification algorithm on second fitting

4.2 Housing Inference and Prediction

4.2.1 Data Cleaning and Processing

A few steps must be taken in order to prepare the complete *housing* dataset for analysis.

First, we make sure that the features that are loaded in are handled/processed as the right format. All of the data is originally imported and set as numeric features. Given our somewhat limited range of data, there are only two features that need conversion:*D-rateincrease* and *D-ratedecrease* . These are converted into categorical features, where 0 signals an absence of change in interest in the last 30 days, and 1 the opposite.

Next, missing data in the dataset is handled. This is referred to as *NA* for "Not Available". In the dataset the only columns containing missing values are the *BRA* and *Byggeår* columns, with 0.006 and 0.0045 percent missing, respectively. There are several ways of handling NA´s. We choose to remove NA´s, as the amount of data lost through the removal of NA´s does not amount to a considerable total loss of data, with over 8000 rows of data lost. The process of handling NA´s and feature processing have been done in both R and Stata to accommodate studies done on both platforms.

4.2.2 Partitioning and Resampling

The housing set is split up into a training set containing 80 percent of the initial dataset and a test set making up the remaining 20 percent.

With xgboost, a k-fold cross validation method will be used. K-fold cross validation is an approach based around randomly dividing a set of observations into k groups, or folds, of approximately equal size (James et.al, 2013). Each fold is then treated as a validation set, with the remaining folds being used for fitting the method. The resulting prediction error from each iteration of the method is computed and assigned to the one held-out fold. The procedure is repeated k times, shifting the role of held-out fold onto a new fold. The resulting number of prediction error-values from the different folds will be used to compute an overall model prediction error by averaging the different values.

When choosing the amount of folds to divide the data by, there are some trade-offs that should be considered. Firstly, k= 5 or k= 10 has been shown to be optimal when it comes to minimization of prediction error, where larger k-values typically produce worse mean prediction-values . Secondly, the bias-variance trade-off is also a factor in k-fold CV, where lower k-values typically result in a more biased model, while larger k-values increase the amount of variance present in the model. In addition, an increasement of folds also increase how computationally intensive the fitting process is. Based on this, we choose to set k=5.

For linear regression prediction, we will use leave one out cross validation (LOOCV). This method is similar to k-fold CV, but the amount of folds used to compute a model prediction error equals the amount of rows in the dataset n-1. As stated earlier, having

a large amount of folds can significantly increase computational intensiveness. However, LOOCV is optimized for least squares linear regression, enabling the computational cost of the method to be equaled to one single model fit (James et.al, 2013).

4.2.3 Model Performance Metrics

We will base the measuring of performance differently based on if the method is inference or prediction.

When discussing the findings by inference, coefficients of variables that has been deemed significant on a 5 percent level will be discussed. Adjusted r-squared levels are also deemed important. This metric shows how much of the variation of the response variable the explanatory variables account for. The r-squared metric is measured on a scale from 0 to 1, where a model with an r-squared of 1 is perfect in explaining variation of the response variable, while 0 explain no variance at all.

For prediction, accuracy will be measured in terms of RMSE. RMSE shows the standard deviation of the residuals; a measure of error in terms of distance between the regression line and the different data points. The better the regression line fits the data points, the lower the RMSE-value will be. Lower is better. RMSE-values range from 0 to infinity, and scales with the response variable. A response variable listing housing prices will as a result likely have a much larger RMSE-value than one listing car prices. RMSE is computed for the training and test data, enabling both baseline evaluation and out-of-sample performance.

$$RMSE = \sqrt{\Sigma(P_i - O_i)^2/n}$$

Here, P_i is the predicted value for the i^{th} observation in the dataset. O_i is the observed value for the i^{th} observation in the dataset. n is the sample size.

4.3 Hyperparameter Tuning

Tuning our recurrent neural network hyperparameters, we use functionality found in both the keras package and the caret package for tuning of our neural network and the xgboost algorithm, respectively. Both packages enable hypertuning through random search and grid search. Grid search systematically fits a model for each of the combinations of hyperparameters that are set for tuning(Brownlee, 2019). A downside with the grid search method is also a part of its strength. The thoroughness of the model guarantees the best model performance within the parameters that are set for tuning. However, this is also very computationally demanding, and computation time increases exponentially with each new option added for tuning.

This is why random search is employed. Random search enables the fitting of random combinations of hyperparameters, and while the method is not as thorough as grid search, it has been shown to be more efficient (Brownlee, 2019). We choose to try only 5 percent of the total number of hyperparameter combinations in our neural network, due to the large amount of hyperparameters that need tuning. For the xgboost model we employ a non-random grid search. The tuning parameters that are employed for RNN-classification and XgBoost regression can be seen in Appendix .

5 Analysis

In this chapter we will present the different models used in our analysis, and the results they yielded. The importance and relevance of these results will be discussed thoroughly in chapter 6.

For our analysis, we use two different approaches. First, we analyze our data using regular OLS regressions, and look for significant relationships in the dataset. Second, we use machine learning methods in XgBoost to study the effect of sentiment on predictions.

5.1 Inference Analysis

Sentiment and price

The first thing we do is to investigate whether there are any basic relationships between our sentiment variable and the behavior of buyers and sellers in the real estate market. We run separate regressions on all our different time lags, and the results are displayed in table 5.1 below.

	(1) Pris	(2) Pris	(3) Pris	(4) Pris	(5) Pris
SentimentD~t	114018.7***				
	(13.28)				
Sen~1Justert		147833.9***			
		(16.21)			
Sen~7Justert			284275.2***		
			(21.55)		
Se~30Justert				326294.9***	
				(21.76)	
Se~90Justert					329204.6***
					(20.01)
_cons	4833956.2***	4796174.3***	4652679.9***	4625751.7***	4658246.9***
	(298.57)	(285.43)	(245.07)	(225.06)	(212.66)
N	90706	89143	94019	92650	89238
R2	0.19%	0.29%	0.49%	0.51%	0.45%

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 5.1: Different sentiment lags and price

Running a number of very simple regressions with the price as the dependent variable,

and the sentiment scores with five different lags as explanatory variable, we see that one unit of extra positive sentiment is associated with an increase in the sales price of between 114,000 and 330,000 Norwegian kroner, depending on which lag that is applied.

The coefficients in all the regressions have p-values of zero, so we can be confident that an increase in sentiment score in fact is associated with higher prices. The R-squared values are at most 0.51 percent, indicating that sentiment alone only explains a tiny fraction of the overall variation in price.

This is not very surprising, as the price of a random house is likely to be determined more by its individual qualities rather than sentiment. However, by aggregating housing prices on the monthly level, individual differences in qualities are no longer a factor. Thus, we would expect sentiment to have higher explanatory power.

Sentiment and the housing price index

In the next model, we switch to the housing price index for Oslo as our response variable. We again run separate regressions on all our different time lags, and the results are displayed in table 5.2 below.

	(1) BPIndexO	(2) BPIndexO	(3) BPIndexO	(4) BPIndexO	(5) BPIndexO
SentimentD~t	6.935***				
	(124.38)				
Sen~1Justert		7.041***			
		(118.92)			
Sen~7Justert			15.64***		
			(208.75)		
Se~30Justert				19.40***	
				(257.00)	
Se~90Justert					20.79***
					(294.90)
_cons	295.4***	295.5***	286.4***	283.1***	283.5***
	(2808.99)	(2709.22)	(2657.39)	(2736.21)	(3020.35)
N	90706	89143	94019	92650	89238
R2	14.57%	13.69%.	31.67%	41.62%	49.36%

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 5.2: Different sentiment lags and the housing price index

We observe that the R-squared values increase up to 49.36 percent for the 90-day moving

average, meaning that the change in sentiment explains almost half of the variation of the index. The coefficient for the 90-day moving average indicates that one unit more of measured sentiment is associated with a jump of over 20 points in the index, or about one fifth of the increase in the index from the beginning of 2016 through 2020. The P-value is still zero for all the different lags, thus we can be sure that the correlation is in fact positive.

These results show that information provided in the news articles is very much associated with movements in price. However, based on such a simple regression alone, we cannot say anything about whether people are responding to the fundamental information provided in the articles, or whether *sentiment* plays a role.

Sentiment and price - with control variables

We further try to separate the effects of events and circumstance from the effect of sentiment. A way this could be done is to introduce control variables for the events believed to affect the behavior. The idea is that the effects caused by the events are captured by these control variables. This leaves the coefficient of our independent variable *sentiment* to be as unbiased as possible. Based on the discussion on variables in chapter 3, we run several multiple regressions, where the output is shown in table 5.3.

All the regressions in the last series have got an explanatory power (R-squared) of a little over 66 percent, meaning that the variables included in the regressions account for about two thirds of the price variation in the data-set, and we see an effect from the different sentiment lags on price from around 19,500 to almost 98,000.

The only control variable not significant at the 5 percent level in any of the regressions is the unemployment variable, while monthly GDP change is only significant at the 10 percent level in the regressions with the 90-day rolling average lag.

We see that the coefficient for *sentiment* is highly significant for all the regressions, and the coefficient grows successively as the lag period increases. It is also interesting to note that even sentiment data from the same day as the transaction occurs, is associated with changes in the final price. How these outputs could be interpreted and what they mean in an economic context will be discussed in chapter 6.

Sentiment and price difference

	(1) Pris	(2) Pris	(3) Pris	(4) Pris	(5) Pris
SentimentD~t	19562.2***				
	(3.56)				
BRA	48273.0***	48260.1***	48367.6***	48493.3***	48852.9***
	(417.12)	(415.54)	(423.69)	(421.11)	(415.17)
Byggeår	-3082.7***	-3145.2***	-3119.7***	-3131.9***	-3187.3***
	(-23.82)	(-23.96)	(-24.37)	(-24.28)	(-24.00)
D_renteopp~g	-86296.9***	-87912.4***	-83469.9***	-88733.0***	-71989.2**
	(-3.59)	(-3.62)	(-3.59)	(-3.81)	(-3.04)
D_rentened~g	-191647.9***	-195981.6***	-173624.3***	-180061.7***	-110465.6**
	(-6.44)	(-6.67)	(-5.89)	(-6.04)	(-2.92)
Styringsre~e	-105650.5***	-107056.9***	-116866.5***	-100960.4***	-96897.5***
	(-6.17)	(-6.22)	(-6.84)	(-5.84)	(-5.53)
Navledighet	-0.0936	-0.225	-2.110	-1.447	-1.986
	(-0.08)	(-0.19)	(-1.76)	(-1.15)	(-1.44)
BNP	18745.7**	19050.3**	12517.9*	14011.3*	11351.1
	(3.14)	(3.19)	(2.11)	(2.31)	(1.81)
IndexKvartal	50492.8***	49966.9***	48308.5***	45937.9***	41880.1***
	(40.56)	(39.93)	(37.55)	(33.77)	(28.76)
Sen~1Justert		28218.7***			
		(4.87)			
Sen~7Justert			71243.7***		
			(7.29)		
Se~30Justert				81665.8***	
				(6.71)	
Se~90Justert					97811.9***
					(6.81)
_cons	6805962.3***	6930003.4***	6872651.8***	6888715.3***	7009832.9***
	(26.89)	(27.00)	(27.46)	(27.31)	(26.98)
N	89964	88418	93260	91898	88523
R2	66.36%.	66.59%	66.25%	66.26&	66.39%

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 5.3: Sentiment and price - with control variables

Even though the effect of sentiment on nominal price is the main focus of this study, our data-set also contains information on the difference between the asking price and the final price. To see if there are any related trends, we run a series of regressions where we replace the dependent variable price with price-difference. The results are illustrated in appendix A0.1, and are heteroskedasticity robust.

We see that all the different versions of the sentiment coefficient are highly significant, and have values in the range between around 11,500 up to about 96,500. The control variables are also highly significant, with the exception of property size. This is not surprising, since we do not see any logical reason why real estate agents and sellers capabilites of reading the market potential would vary for homes of different sizes. When it comes to the R2, we see that the variables included in the model explains between 3 and 3.5 percent of the variation in price difference. This is expected, as a large gap between asking price and final price usually comes from mispricing or that several bidders are very eager to buy a property and keeps overbidding each other. These are factors that we do not have data on in our dataset.

5.1.1 A Discussion on Robustness

Before discussing the implications of the output from the regressions, we need to know whether the results can be trusted or not. Specifically, we are interested in whether the sentiment coefficient in the regressions is biased or not. According to the Gauss-Markov theorem, there are four conditions that must be satisfied in order to conclude that we have an unbiased estimator (Woolridge, 2014). The model must be (1) linear in parameters, (2) randomly sampled, (3) have no perfect collinearity amongst the independent variables, (4) have zero conditional mean. If these four conditions are satisfied our model is generally unbiased, and if a fifth condition – homoskedasticity – is also satisfied, our estimator is the best linear unbiased estimator (BLUE).

Linear in parameters

To get an unbiased coefficient using linear regression, the relationship between the dependent and independent variables must be linear to begin with. We have generated a number of scatter plots that illustrates the trends between our dependent variable and the independent variables. They show that relationships are indeed linear, and can be

studied in appendix A0.3.

Random sample

Another condition is that the sample must be randomly drawn. The intuition is that in order to say something about an entire population, the data must be representative. An example of the opposite would be to try to figure out the average IQ of the population in a country, but only test university students. We have data from *all* transactions in Oslo for a long period of time, thus we can be sure that our data is representative: we have examined the whole population in Oslo, not just a sample.

To say something about whether our results are valid for all of Norway is a little more difficult to do, but when we create a scatterplot with sentiment and the nationwide housing index shown in A0.5, we see a clear linear relationship there too. This implies that relationships and overall trends will also be present on a national level, but the coefficients might have different values.

No perfect collinearity

In our model, none of the independent variables can be a constant, and there cannot be any perfect linear relationships among the independent variables. If a perfect linear relationship is present, it is not possible to distinguish between their effects, and the model is biased. We know that we have no constant variables, and with the use of correlations matrixes A0.4 we observe that none of the variables are perfectly correlated.

Zero conditional mean

The zero conditional mean condition says that the error term of the regression equation should have an expected value of zero for any value of the independent variables. This means that the positive and negative deviations of our observed values from the model should even each other out, on average. In our case this could mean that a transaction with a higher price than the model predicts will be cancelled out by a transaction with a lower price than predicted. A way to check for this is to plot the fitted values along with the residuals in a scatter plot, and see whether the pattern is symmetrical. In our case the pattern is fairly symmetric with the exception of a few outliers, and we can assume that the condition is satisfied. The scatter plot is found in appendix A0.6.

No heteroskedasticity

Heteroskedasticity is when the error term of the regression equation has different variance for different values of our independent variables. To test for this, we ran a Breusch-Pagan / Cook-Weisberg test for all the versions of the regression. The tests yielded a p-value of 0, which means that we must reject the null hypothesis of constant variance, meaning we have heteroskedasticity.

Conclusion on robustness

Since condition 1 through 4 is satisfied we can conclude that the model is unbiased and consistent, and we can continue to discuss the economic relevance of the results. Since we have heteroskedasticity we must however recalculate our standard errors and T-statistic. The reason is that the initial ones were calculated assuming that the variance of the error term was constant for all values of our independent variables. When we know that this is not the case, we must adjust for it. As the standard error is related to the T-statistic – and thus the significance of the coefficients – and the confidence intervals, these values might also change. The fact that we have heteroskedasticity also implies that we cannot rule out whether there are OLS models that are better than ours, and we cannot claim that our model is BLUE.

5.1.2 Linear Regression with Heteroskedasticity-robust Standard Errors

Given the conclusion from the previous chapter, we recalculate the regression using heteroskedasticity-robust standard errors. The output is shown in appendix X.

As expected, none of the coefficients or R2-values change. When it comes to statistical significance, the dummt variable for rate increase is now significant at the 1 percent level instead of the 0.1 percent level in regression 1 and 2. This is still what we would describe as significant results, since we are over 99 percent sure that the coefficient has the correct sign. Thus, our analysis and upcoming discussion will not be affected much by the presence of heteroskedasticity.

5.2 Prediction

5.2.1 Linear Regression

We start by predicting through the use of linear regression. The linear regression will not be able to capture non-linear relationships in the data. We use the linear regression as a benchmark for the performance of the XgBoost model, since we suspect it will give a larger prediction error than a machine learning model.

Table 5.1: LM RMSE with reference and sentiment

	Reference	**Sentiment7**	**Sentiment30**	**Sentiment90**
Train RMSE	1 647 270	1 651 239	1 651 575	1 651 798
Test RMSE	1 687 204	1 692 278	1 692 316	1 692 455

In table 6.1 we present the results from the linear regression model. Both the results from training and test RMSE are shown in the table, with a row showing performance of a reference model, consisting of all features excluding any of the sentiment variables. From left to right we showcase the performance achieved by the model with three different lags: 7 days, 30 days and 90 days. These lags were chosen to showcase the impact of sentiment on three different types of buy and sell decisions, where the 7 day lag represents a more spontaneous transactions than what a 90 day transaction would be. The variables are somewhat correlated, as the shorter lags are subsets of the longer lags.

From the table we observe that the reference model does a better job at predicting than when sentiment is introduced. With an increase of 5000 RMSE from reference to Sentiment7, the model performance is lessened by 0.3 percent. We see an almost identical decrease in model accuracy between the reference model and the other lagged variables.

5.2.2 XgBoost

We repeat the process shown with linear regression.

Table 5.2: XgBoost RMSE with reference and sentiment

	Reference	**Sentiment7**	**Sentiment30**	**Sentiment90**
Train RMSE	1 394 200	1 433 102	1 384 691	1 411 897
Test RMSE	1 434 680	1 471 292	1 464 083	1 398 504

In table 6.2 we now see the results from prediction done by xgboost over a 90 day horizon. As can be seen from the overall decrease in RMSE-values, the xgboost model provides a more accurate model fit compared to linear regression. Comparing test RMSE of both LM and XgBoost reference models, the machine learning algorithm increases accuracy by 14.96 percent. This indicates that the ability of xgboost learning algorithms are better suited to capture a relationship between housing price and its predictor variables. This could be an indication that certain data exhibits non-linear relationships that a linear regression model is not as effective in capturing.

Looking at the developments in test RMSE, the introduction of sentiment7 reduces accuracy by 2.55 percent from the reference model. Sentiment30 is not much better with a reduction in performance by 2.04 percent. Sentiment90 is the only variable of the three sentiment-variables that demonstrates an increase in accuracy with a 2.52 percent reduction in RMSE compared to the reference model.

Also, comparing the relative difference in training RMSE and test RMSE, none of the models show signs of underfitting or overfitting. This is a sign that the xgboost model does a good job of generalizing onto the test set.

5.2.3 Variable Importance

As previously mentioned, it is practically impossible to accurately capture the form of $\hat{f}$ when a machine learning model has been used. ML-models give little insight into how the function works, but some output that can be captured is the importance of variables. This is done through the *varImp* function in R. The function displays the relative importance of variables used in the data. As can be seen from figure 6.4, BRA is measured highest, with an importance of 100. This is a relative scale, meaning that BRA with a value of 100 is almost 8 times as important in the prediction of housing prices as the next most

important variable, Byggeår. Here we see a difference occurring between the reference and the model containing sentiment90.

Also, in the variable importance plot we have imported three different sentiment variables from their respective three data sets, in order to illustrate the difference in variable importance. As can be seen from the plot, SentimentL90 is more important to the prediction than SentimentL30, which is more important than SentimentL7.

We also see that most of the macro variables have little to no predictive power relative to the housing specific variables, and even sentiment.

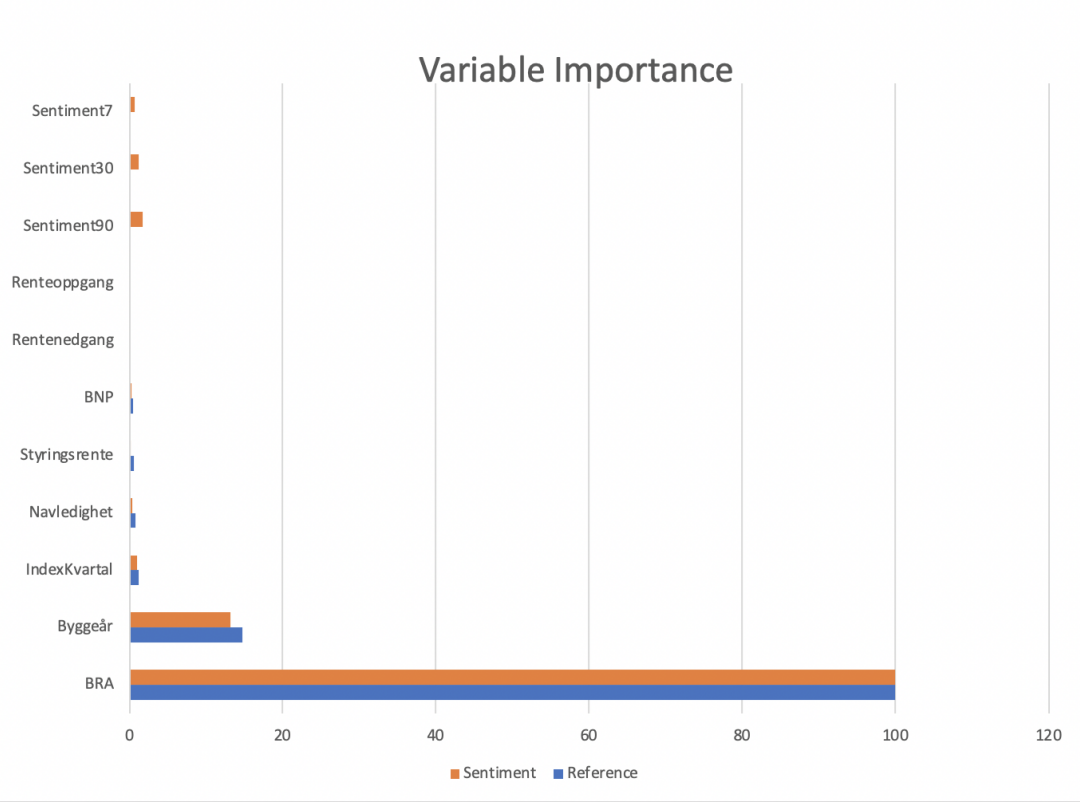

Figure 5.4: Recurrent Neural Network

6 Discussion

In this chapter we will discuss the economic significance and plausibility of the results from chapter 5, discuss potential applications of the results, and discuss some limitations while we point to some areas of interest for future research.

6.1 Implication of Findings

In the first and second regression series displayed in table 5.1 and table 5.2, we observe that the sentiment variable is associated with higher nominal prices and higher values of the housing index. It useful information to know that these things are correlated, and the implication is that buyers and sellers in the real estate market respond to information, meaning that price development does not simply follow a random walk. Quite the contrary actually, as the variation of the sentiment variable explains almost half of the variation in the housing price index.

This could suggest that newspapers are an important and trustworthy source of information, but since the relationships are measured on a pure correlational basis, we are unable to draw any certain conclusions. It is possible that information acquired elsewhere is much more important, but it is hard to imagine for example the website of the central bank competing in reach with the mass media. Therefore, we believe that our data suggest – but does not prove – that people respond to information provided in the media, which again implies that they trust the information provided.

We are, however, more interested in how *actual sentiment* – not the value of our sentiment variables - is affecting prices, and to do that we must control for the fundamental events and information that are likely to correlate with our sentiment variables. This is what we are aiming to do in the next step.

From the first regression in table 5.3, we saw that one extra unit of measured sentiment on the same day as the transaction is associated with a higher sales price at around NOK 19,500. It may seem odd that newspaper article from the same day has potential to influence the transaction price that very day, but real estate purchases in Norway usually takes the form of an auction. It is not uncommon that several bidders compete all through

the morning and the early afternoon for a property they favor. It seems perfectly plausible that positive newspaper articles consumed while participating in an auction may lead to a more optimistic mindset, and higher marginal willingness to pay.

From the second regression in table 5.3 we that one extra unit of sentiment measured the day before the transaction is associated with a price increase of about NOK 28,000. Usually, potential bidders physically view and inspect the property the day before the bidding round, and a buyer will typically make up their mind on whether they are interested or not that evening. It is a reasonable assumption that many bidders also decide on a maximum price that afternoon or evening, and it seems perfectly plausible to us that positive news consumed that day may nudge someone to agree on a higher limit.

From the third regression in table 5.3, we saw that one extra unit of sentiment measured as the average of the 7 days prior to the transaction date is associated with a price increase of about NOK71,000. In most cases, the date and time of the viewing/open house of the selling property is announced about a week beforehand. This announcement is typically stated in the online advertisement of the property, along with property information and a number of photographs. It seems perfectly plausible that exposure to several positive newspaper articles in the period between one first discovers an interesting property, and the day where you get to see the property physically might lead to a higher willingness to pay. Buying a home is a big decision, and it is reasonable to assume that most buyers are likely to discuss the purchase decision with friends and family, and positive news articles may bring extra optimism into such discussions.

In the fourth, and fifth regressions in table 5.3 we saw that one extra unit of sentiment measured as the average of the 30 and 90 days prior to the transaction date is associated with a price increase of almost NOK82,000 and NOK98,000, respectively. This tells us that positive newspaper articles in the period long before the property is even listed, is associated with higher prices. Also in this case, the relationship seems plausible. Many households probably realize quite some time in advance – perhaps several months or even quarters - that they are eager to move, for example if they are expecting children. It is perfectly plausible that exposure to positive news articles in the period where potential buyers are actively considering making a move on the real estate market could lead to higher willingness to pay.

To assess whether the regressions as a whole make sense and seems plausible, it is natural to discuss the coefficients for the control variable as well. We see that one square meter of extra space is associated with about a higher price of over NOK 48,000. This makes sense, as larger properties are expected to sell for more, all else equal. We see that newer buildings sell for less than older ones, which might be a surprise. All else equal, we would expect newer buildings to sell for more, but since we have not controlled for location within Oslo, we believe that the coefficient captures some of the effect related to that the buildings in central and more exclusive parts of the city tend to be much older than the buildings in the less central districts.

Further, we observe that an increase in the key policy rate withing the last 30 days prior to the transaction is associated with a price drop of between NOK 72,000 and NOK 89,000. This is expected, as higher interest rates make it more expensive with a mortgage and reduces the supply of capital from banks to potential buyers. More surprising is it that a decrease in interest rates the last 30 days is associated with an even bigger price drop, at between NOK 110,000 and NOK 191,000. We were expecting the opposite to happen, as the logic applied to the increasing rate would be applied in reverse. However, after some consideration we are fairly sure that the relationship is reasonable, since the lowering of the key policy rate usually is a response to an external shock such as a significant drop in the oil price or a pandemic. Since the shock and the lowering of the rate is so closely connected in time, we find it reasonable that prices are lower after such an announcement by the central bank.

For the general level of the interest rate, however, we observe that higher levels are associated with lower prices, which we expect. We also observe lower prices when the unemployment is higher. This is expected, but the coefficients are not statistically significant. We also observe that the positive GDP-growth is associated with higher prices, which we also find reasonable, although the coefficient is not significant at the 5 percent level for the 90-day sentiment lag. Lastly, we see that for each quarter passing, the price increases successively, which is expected due to regular inflation trends. Overall, we see the effects that we initially would expect, except for the rate decrease announcements which still found reasonable after some thinking. This means that we do not have any red flags in our models, and are strengthened in our believes that our sentiment coefficients

have captured some true effects.

As we have seen, positive sentiment is associated with higher prices in the longer run, as well as in the shorter run, all the way down to the same day as the transaction. In our opinion, these relationships are reasonable, as we believe consumption of positive news may influence one's level of optimism on the longer run, as well as providing a "nudge" during a heated situation such as an auction.

We believe a typical newspaper article related to the housing market is something that contains two components. First, it contains actual information about something fundamental that has happened, such as a change in the interest rate. Second, it contains information such as pictures, graphics, headlines, quotes, and other information that goes beyond simply describing a fundamental event. Our sentiment scores that we use in our regressions are functions of both components, but when we control for these fundamental events, we still see that the sentiment coefficient is significant both in statistical and economic terms. This implies that decision makers responds to news articles beyond what they are expected to do, based on the fundamental information alone.

In our opinion, it is reasonable to interpret our results so that higher sentiment – as it has been described and defined in chapter 2 – is associated with higher prices. Our implicit hypothesis throughout this thesis has been that the price is a function of financial considerations, non-financial considerations, *sentiment*, as well as an error term. This could be written formally as:

P(House) = PV(future cash flows) + non-financial utility + sentiment + error term

Based on the regression series in table 5.3, this equation holds up. We see that financial considerations such as interest rate, changes in interest rate, GDP development, and inflation are significant factors in determining the price. Further, non-financial considerations such as size and construction year also play their part. One could argue that these factors are also indirectly part financial, as bigger houses is likely to sell for more in the future, but that is less important here. Much more important is that we have found the presence of a sentiment part of the equation, and we have found it to be statistically significant as well as large enough to matter substantially in economic terms.

This is also backed up by the regression series in appendix A0.2, which shows a relationship

between sentiment and the difference between asking price and final price. It was expected that such an effect would be present in the short term, as the "mood" might shift between when the asking price was decided, and the transaction day. A little more surprising is it that the effect is (1) present, and (2) much stronger in our regressions with the long-term sentiment. However, it makes little sense that real estate agents should be able to fine-tune their price estimates in real time. The past transactions they use as their benchmark might be conducted months ago, and many real estate agents probably want to estimate the price conservatively in order to avoid angry customers.

We believe that these results are evidence that the housing market is in fact not perfectly efficient. The reason is that we observe that the players in the market "overreact". They respond to the information provided in the newspaper articles, but if the information is "wrapped" in positive or negative wordings they respond more than the fundamental information suggests that they should.

These findings are also consistent with the results from the prediction analysis, where the XG-Boost machine learning model finds that adding the 90-day sentiment lag - the strongest coefficient from the linear regression - improves the predictive power of the model.

Whether the results are valid for Norway as a whole, rather than just Oslo, is another question. The sentiment data is retrieved from national newspapers, and we do observe a correlation between those scores and the national housing price index. In addition, one would intuitively expect people to have the same psychological mechanisms regardless of whether they live in Oslo or not. This points towards a more general applicability. Still, Oslo have the highest housing prices in the country, and have also seen the highest growth in prices the recent years. Such factors could induce a special dynamic with more speculative behavior, and our analysis is after all concentrated about transaction data from Oslo. We believe there is reason to suspect the same the same tendencies nationwide as in Oslo, but in the end we only have conclusive empirical evidence for Oslo.

6.2 Financial and Economic Applications

If your findings are in fact true, they could potentially have significant real life-impact on the decisions of home buyers as well as financial investors if they respond rationally

to the insights. Based on the sentiment measured and analyzed from newspaper articles, buyers and sellers could time their behavior to maximize their returns. Any economist could easily imagine the implications if there was a reliable tool available to predict if the stock market would perform relatively good or relatively bad for the coming periods: abnormal returns for those who had access to the tool, and in some cases massive arbitrage opportunities.

However, several things point towards the applications being a bit more limited. First, transaction costs such as stamp duty, real estate agents, and actually switching homes makes it hard to utilize insights on price trends, especially if they are not very large. Second, it is fair to assume that most people buy and sell basically at the same time, as they need to have a place to live. This fact deprives them of the opportunity to time their behavior based on sentiment analysis. Third, the housing market lacks instruments such as derivatives and the option to short assets, implying that an investor cannot take advantage of mispricing that easily.

Still, for real estate investors that are planning to liquidate parts of their portfolio, but have the flexibility to decide exactly when they will do it, these findings might be helpful. The same applies to anyone that has some sort of flexibility on when to buy or sell, and have to deal with the transaction costs either way.

6.3 A Discussion on Causality

With findings of such strong statistical and economic significance, it might be interesting to discuss whether the relationship between sentiment and price is a causal one, that is that higher sentiment causes higher prices. Based on the data we have and the methods we have chosen, we are unable to say anything certain about this. Ideally, we would have wanted a randomized controlled study or another type of experiment, but in real life you cannot deprive half the population of access to news media, it is simply a question of practicalities.

Although we are unable to conclude for sure on causal relationships, we find it appropriate to discuss some aspects that may say something of whether it is likely that a causal relationship might be present. We use the Bradford-Hill criterions for causality as our baseline for the discussion (Schünemann, Hill, Guyatt, 2011). Although the criterions

initially were developed for use in medicine, most of the criterions are applicable also in other disciplines.

We would like to point out a few factors which leads us to think that there might be a causal relationship. First, it does not seem reasonable that the effect is caused by reverse causality, that is that prices affect sentiment. The reason is that we use sentiment variables collected before the prices are formed, prohibiting prices to influence sentiment. Second, the relationships are strong. All the sentiment coefficients in all the regression are statistically significant with very low p-values, and the coefficients are also economically significant: they are important also in the real world. Third, we find a causal relationship to be plausible, as we discussed in the previous section. It simply makes sense that optimism on a societal level "transmits" to the individual home buyers. We do not claim that these factors in any way prove a causal relationship, but we believe that pointing out these simple facts are appropriate.

6.4 Limitations

Although we are confident in our methodology and our analysis, there are some limitations present. One limitation that seems intuitive at first, is that we seem to have a circularity problem: we do not know if newspaper articles shapes sentiment, or is a reflection of it. However, we do not really care about this in our analysis. All we care about is whether sentiment affects prices, the root cause of the sentiment is not important. Thus, other limitations are more prevalent.

First, it is unlikely that we have been able to fully control for all the relevant fundamentals. We can assume that our sentiment coefficients consist of one part "true sentiment" and one part rational response to information provided. If we have controlled for the most substantial fundamentals, the part consisting of "true sentiment" will be relatively large. If we omit some important fundamentals, the part consisting of "true sentiment" will be lower. Truth be told, we do not know for sure how much of the coefficient's effect that are sentiment, and how much are related to fundamentals we have failed to control for. Thus, we are fairly sure that our model has some degree of omitted variable bias. We have, however, controlled for the same things (and more) that have been controlled for in the related and well-cited studies that are reviewed in chapter 2.

A second limitation is introduced in the training of our RNN classification model. Here, the data set is unevenly scored on a scale from 1-6, where the amount of data labelled 1,2 and 6 is small relative to the other 3 groups. This could have caused our RNN classification algorithm to give a worse performance on very negative or very positive language, - language that would otherwise be scored at 1, 2 and 6.

Third, the data received from Eiendomsverdi is definitely limited in terms of how many features are provided. As an example, we know location to be a strong predictor of housing price, and we know that Eiendomsverdi has access to this type of information. The results from our analysis will without question reflect the lack of features we were provided with.

Fourth, when doing prediction with neural networks, a large amount of data provided is necessary. Seeing as we chose to limit our analysis to a span of five years, the amount of transactions happening in Oslo within this time frame is naturally a lower amount than if we were to extend the time frame. The DFNN performance is likely not optimized on the limited amount of data that is present in the housing data set.

Fifth, the pre-trained word embeddings are exclusively fast text-skip grams. While we optimally would want a diverse set of NLP inputs in order to optimize performance, natural language processing resources in the Norwegian language are extremely limited relative to the resources available in English. This might have reduced the performance of our classifier to some degree.

6.5 Further Research

When it comes to further research, we believe that it is reasonable to use our finding as a starting point: sentiment is associated with prices. In interesting approach would in our opinion be to categorize the sentiment scores by newspaper, by type (which fundamentals are being covered), or both. At this point, we are only able to say something about the average effect, since all articles are in one unsorted pool.

A qualitative study of the content of the newspaper articles is also likely to be interesting, as it may uncover that some fundamentals are frequently mentioned, but not controlled for. A consequence would be that the sentiment coefficient would reflect the “true” effect more precisely.

An experimental study where a selected group of individuals were asked to "grade" how positive or negative an article is perceived could also be interesting. One could imagine giving the same fundamental information (such as a change in the interest rate) to two groups, but give the groups articles that differ in how positive or negative they are worded, and investigate any differences.

Lastly, any studies that may shed light on the causal relationships would certainly add something to the table. How such a study might be designed is a question that we leave for others, but one approach might be to utilize a time series analysis with panel data, in order to remove the individual specific errors in the data.

7 Conclusion

In this thesis we have studied the relationship between information and sentiment provided through the news media, and housing prices in Norway. To do this, we have performed sentiment analysis on newspaper articles, and analyzed the relationship with housing prices using an OLS regression model, and machine learning models.

We have found that the sentiment values measured from the news explains almost half of the variation in the housing market index for Oslo, leading us to conclude that people are likely to respond to information provided in newspaper articles. This means that the housing price development cannot be a random walk, and that the housing market must be more efficient than the weak form om the EMH suggests.

Further, we have found that our sentiment coefficient is both statistically and economically significant, after we have controlled for fundamentals in our OLS regression. This leads us to conclude that sentiment – as it has been defined in chapter 2 – is associated with the price development in the real estate market. This means that people respond to events and circumstances to a greater extent than what the fundamental information justifies, which means that the housing market does not fully satisfy the criterion of the semi-strong form of the EMH: that prices reflect all public information.

Appendix

A0.1 Price and sentiment

	(1) Pris	(2) Pris	(3) Pris	(4) Pris	(5) Pris
SentimentD~t	19562.2***				
	(3.49)				
BRA	48273.0***	48260.1***	48367.6***	48493.3***	48852.9***
	(134.01)	(133.47)	(136.39)	(137.15)	(134.43)
Byggeår	-3082.7***	-3145.2***	-3119.7***	-3131.9***	-3187.3***
	(-11.16)	(-11.36)	(-11.49)	(-11.39)	(-11.34)
D_renteopp~g	-86296.9**	-87912.4***	-83469.9***	-88733.0***	-71989.2**
	(-3.29)	(-3.49)	(-3.32)	(-3.55)	(-2.84)
D_rentened~g	-191647.9***	-195981.6***	-173624.3***	-180061.7***	-110465.6**
	(-6.65)	(-6.86)	(-6.07)	(-6.24)	(-2.81)
Styringsre~e	-105650.5***	-107056.9***	-116866.5***	-100960.4***	-96897.5***
	(-5.93)	(-5.95)	(-6.56)	(-5.62)	(-5.34)
Navledighet	-0.0936	-0.225	-2.110	-1.447	-1.986
	(-0.08)	(-0.20)	(-1.80)	(-1.18)	(-1.45)
BNP	18745.7**	19050.3**	12517.9*	14011.3*	11351.1
	(3.14)	(3.19)	(2.10)	(2.33)	(1.82)
IndexKvartal	50492.8***	49966.9***	48308.5***	45937.9***	41880.1***
	(41.74)	(40.58)	(39.01)	(34.79)	(29.81)
Sen~1Justert		28218.7***			
		(4.72)			
Sen~7Justert			71243.7***		
			(6.99)		
Se~30Justert				81665.8***	
				(6.47)	
Se~90Justert					97811.9***
					(6.71)
_cons	6805962.3***	6930003.4***	6872651.8***	6888715.3***	7009832.9***
	(12.71)	(12.92)	(13.06)	(12.92)	(12.86)
N	89964	88418	93260	91898	88523
R2	66.36%.	66.59%	66.25%	66.26&	66.39%

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure A0.1: Price and sentiment - robust version

A0.2 Price Difference and Sentiment

	(1) Prisdiff	(2) Prisdiff	(3) Prisdiff	(4) Prisdiff	(5) Prisdiff
SentimentD~t	28092.6***				
	(17.52)				
BRA	0.607	-22.56	-14.60	-24.23	-51.79
	(0.00)	(-0.16)	(-0.11)	(-0.17)	(-0.36)
Byggeår	-285.8***	-291.6***	-286.0***	-282.6***	-289.9***
	(-5.20)	(-5.20)	(-5.33)	(-5.22)	(-5.22)
D_renteopp~g	-16643.1**	-16861.3**	-17180.1**	-17728.1**	-2279.4
	(-2.65)	(-2.64)	(-2.89)	(-3.00)	(-0.38)
D_rentened~g	-103211.6***	-101160.8***	-87636.9***	-73121.9***	-135725.1***
	(-12.89)	(-12.83)	(-11.14)	(-9.39)	(-13.89)
Styringsre~e	-22526.3***	-24757.3***	-29827.5***	-31703.1***	-29822.3***
	(-5.10)	(-5.58)	(-6.77)	(-7.10)	(-6.68)
Navledighet	4.475***	4.506***	2.886***	1.535***	2.235***
	(15.83)	(15.89)	(10.04)	(5.16)	(7.14)
BNP	15689.5***	15464.2***	10566.9***	6500.7***	7132.8***
	(8.92)	(8.80)	(6.15)	(3.84)	(4.17)
IndexKvartal	-17149.2***	-16808.4***	-18526.8***	-19312.2***	-18234.3***
	(-46.40)	(-44.05)	(-50.21)	(-50.27)	(-42.74)
Sen~1Justert		24851.6***			
		(14.30)			
Sen~7Justert			65158.5***		
			(25.29)		
Se~30Justert				92342.1***	
				(27.66)	
Se~90Justert					96469.8***
					(25.54)
_cons	809816.5***	824960.4***	813188.1***	805990.8***	795156.0***
	(7.85)	(7.83)	(8.08)	(7.95)	(7.65)
N	89964	88418	93260	91898	88523
R2	3.35%	3.22%.	3.5%	3.52%.	3.07%

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure A0.2: Pricedifference and sentiment

A0.3 Scatterplot- Price and Sentiment

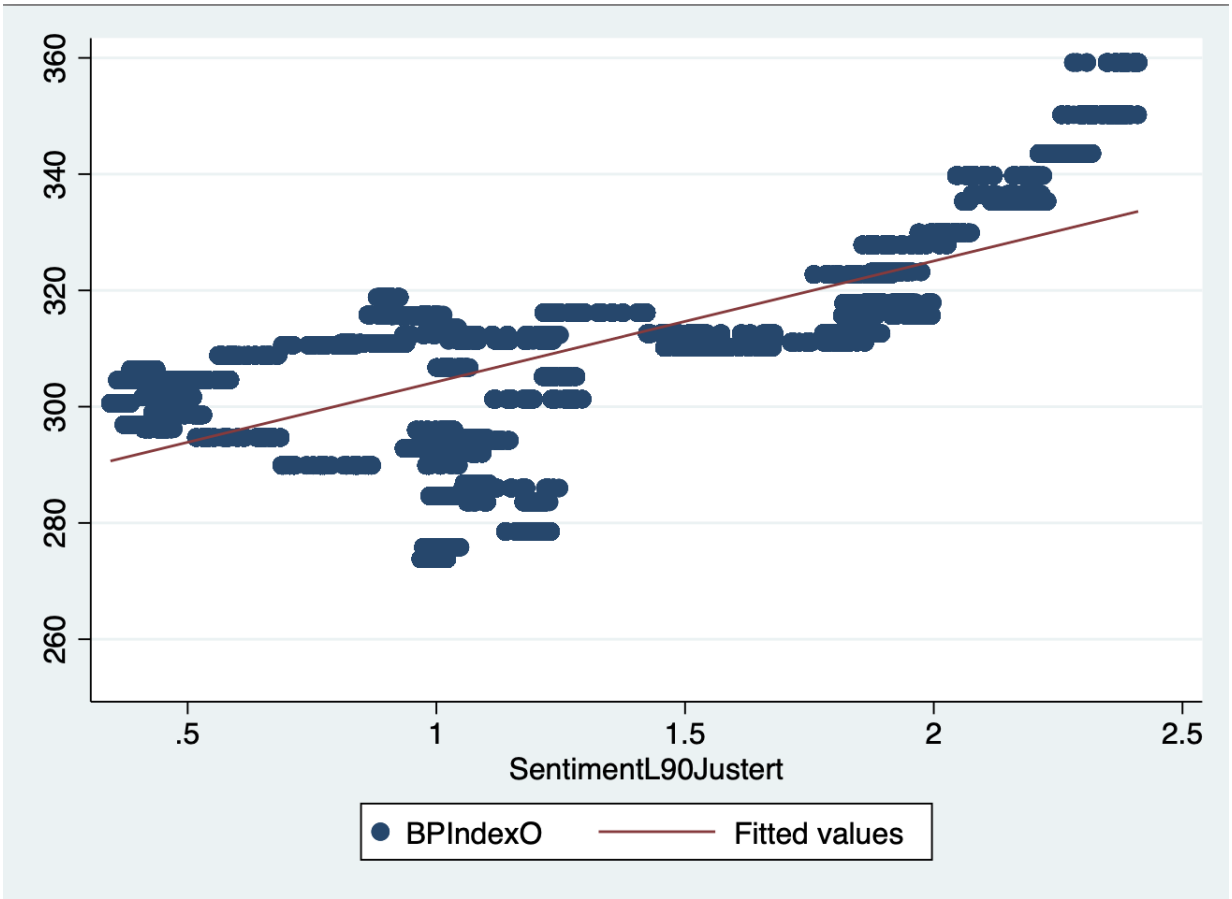

Figure A0.3: Scatterplot showing linearity of price and sentiment

A0.4 Correlation matrix

Correlation matrix:

	Pris	Sentim..	BRA	Byggeår	D_~pgang	D_~dgang	Styrin~e	Navled~t	BNP	IndexK~l
Pris	1.0000									
SentimentD~t	0.0439	1.0000								
BRA	0.8067	0.0015	1.0000							
Byggeår	0.0407	0.0047	0.1046	1.0000						
D_renteopp~g	0.0028	-0.0325	0.0003	0.0057	1.0000					
D_rentened~g	-0.0021	0.1622	-0.0006	0.0034	-0.0638	1.0000				
Styringsre~e	-0.0159	-0.0756	-0.0021	0.0098	0.3396	-0.2002	1.0000			
Navledighet	0.0411	0.3280	-0.0030	0.0087	-0.1276	0.5213	-0.3713	1.0000		
BNP	0.0160	-0.0739	0.0092	-0.0021	0.0494	-0.1790	-0.1597	-0.4387	1.0000	
IndexKvartal	0.1036	0.3629	0.0043	0.0241	0.1420	0.1008	0.0306	0.4533	-0.0371	1.0000

Figure A0.4: Correlation matrix

Scatterplot Nationwide

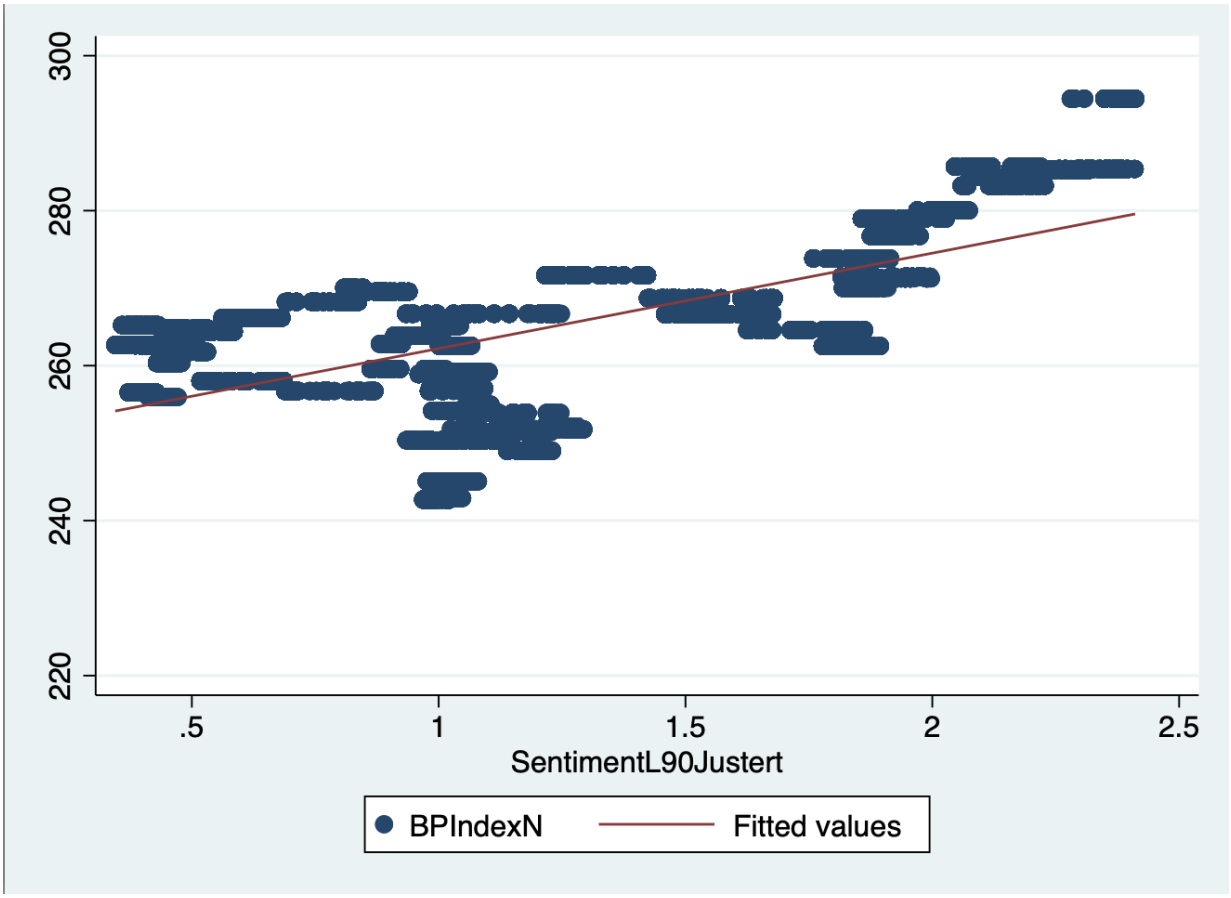

Figure A0.5: Scatterplot showing sentiment and the national housing price index

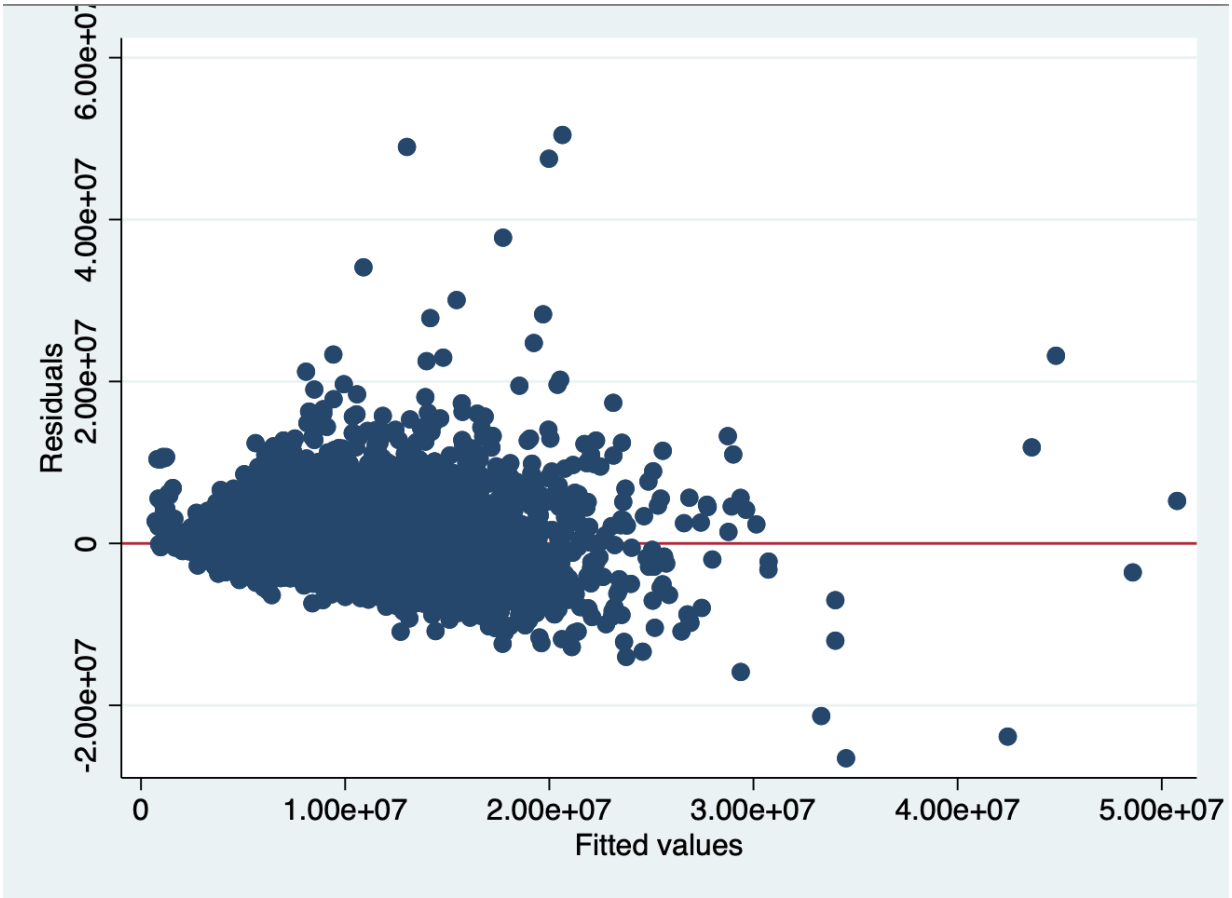

Figure A0.6: Residuals versus fit

A1 XgBoost Hyperparameters

Table A1.1: XgBoost Hyperparameters

Hyperparameter	Value
num-rounds	200
max-depth	4
eta	0.1
subsample	0.7
gamma	0.1
min-child-weight	1
colsample$_bytree$	0.3

A2 RNN Hyperparameters

Table A2.1: Neural Network Hyperparameters

Hyperparameter	Description
Activation Function	RMSProp
Number of Layers	2
Number of Neurons	128, 64, 5
Batch Size	2048
Number of Epochs	500
Learning Rate (LR)	0.0001
Dropout	0.2
L2 Regularization	Not used
Early Stopping Patience	5

References

Fama, E. F., & Laffer, A. B. (1971). Information and Capital Markets. *The Journal of Business*, 289 - 298.

Fama, E. F. (1970). Efficient Capital Markets: A review of theory and empirical work. *The Journal of Finance*, 383-417.

N., D. D., & Berry, M. A. (1995). Overraction, Underreaction, and the Low-P/E Effect. *Financial Analysts Journal*, 21 - 30.

Osland, L., & Thorsen, I. (2008). Effects on Housing Prices of Urtban Attraction and Labor-Market Accessability. *Environment and Planning A: Economy and Space*, 2490 - 2509.

Capozza, D. R., & Seguin, P. J. (1994). Expectations, Efficiency, and Euphoria in the Housing Market. *Working Paper*.

Pollakowski, H. O., & Ray, T. S. (1997). Housing Price Diffusion Patterns at Different Aggregation Levels: An Examination of Housing Market Efficiency. *Journal of Housing Research*, 107 - 124.

Gu, B., & Hitt, L. (2001). Transaction Costs and Market Efficiency . *ICIS 2001 Proceedings*.

Walker, C. B. (2014). Housing Booms and Media Coverage. *Applied Economics*, 3954-3967.

Brueckner, J., Calem, P., & Nakamura, L. (2012). Subprime mortgages and the housing bubble. *Journal of Urban Economics*.

Regjeringen. (2021, 10 22). Retrieved from Regjeringen.no: https://www.regjeringen.no/no/tema/okonomi-og-budsjett/finansmarkedene/boliglansforskriften-1.-januar-202031.-desember-2020/id2679449/

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

Soo, C. K. (2018). Quantifying Sentiment with News Media across Local Housing Markets. *The Review of Financial Studies*, 3689-3719.

Beracha, E., Lang, M., & Hausler, J. (2019). On the Relationship between Market Sentiment and Commercial Real Estate Performance - A Textual Analysis Examination. *Journal of Real Estate Research*, 605-638.

Kirkeby, S. J., & Larsen, V. H. (2021). *House price prediction using daily news data.* Oslo: Norges Bank.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 54(4),* (pp. 82-89).

Chowdhury, G. G. (2003). Natural Language Processing. *Annual Review of Information Schience and Technology*, 51-89.

Universitetet i Bergen. (2021, 11 8). *Opplagstall norske aviser - resultat*. Retrieved from medienorge: https://www.medienorge.uib.no/statistikk/medium/avis/190

Hauger, K. K. (2019, 10 15). Retrieved from Kampanje.com: https://kampanje.com/medier/2019/10/ferske-lesertall-her-er-de-storste-avisene/

Norges Bank. (2021, 11 10). *The policy rate*. Retrieved from Norges Bank: https://www.norges-bank.no/en/topics/Monetary-policy/Policy-rate/

Norges Bank. (2021, 11 10). *Changes in the policy rate*. Retrieved from Norges Bank: https://www.norges-bank.no/en/topics/Monetary-policy/Policy-rate/Key-policy-rate-Monetary-policy-meetings-and-changes-in-the-policy-rate/

NAV. (2021, 11 11). *Arkiv - Helt ledige*. Retrieved from NAV: https://www.nav.no/no/nav-og-samfunn/statistikk/arbeidssokere-og-stillinger-statistikk/helt-ledige/arkiv-helt-ledige_kap

SSB. (2021, 11 11). *Arbeidskraftundersøkelsen*. Retrieved from SSB.no: https://www.ssb.no/statbank/table/08518/

Woolridge, J. (2014). *Introduction to Econometrics.* Cengage Learning.

Schünemann, H., Hill, S., & Guyatt, G. (2011). The GRADE approach and Bradford Hill's criteria for causation. *Journal of Epidemiology & Community Health*, 392-395.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money.* Palgrave Macmillan.

Akerlof, G. A., & Shiller, R. J. (2009). *Animal Spirits.* Princeton University Press.

Soo, C. K. (2015). Quantifying Animal Spirits: News Media and sentiment in the Housing Market. *Working Paper - Ross School of Business*.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis alogrithms and applications: A survey. *Ain Shams Engineering Journal*, 1093-1113.

The International Energy Agency. (2018). *Global EV Outlook 2018, Towards cross-modal electrification.* The International Energy Agency.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* Springer.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Pearson, E. S. (1931). The test of significance for the correlation coefficient. *Journal of the American Statistical Association*, 128 - 134.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

West, M. (2021, 12 15). *Bouvet*. Retrieved from Explaining recurrent neural networks: https://www.bouvet.no/bouvet-deler/explaining-recurrent-neural-networks

Brownlee, J. (2021, 12 15). *A Gentle Introduction to Backpropagation Through Time*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/gentle-introduction-backpropagation-time/

Sinha, N. (2021, 12 15). Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-lstm-and-its-quick-implementation-in-keras-for-sentiment-analysis-af410fd85b47

Github. (2021, 12 15). *Feedforward Deep Learning Models*. Retrieved from Github: http://uc-r.github.io/feedforward_DNN#ff

Haile, T. (2021, 12 15). *What you think you know about the web is wrong*. Retrieved from Time: https://time.com/12933/what-you-think-you-know-about-the-web-is-wrong/

Dingcheng, L. (2021, 12 15). Retrieved from ResearchGate: https://www.researchgate.net/figure/2D-PCA-projection-of-word-embeddings-Five-different-word-clusters-are-shown_fig2_332892222

McCormick, C. (2021, 12 15). *Word 2 Vec Tutorial - The Skip-Gram Model*. Retrieved from mccormickml: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Almeida, F., & Xexéo, G. (2018). *Word Embeddings: A survey.* Rio de Janeiro: Federal University of Rio de Janeiro.

Sarkar, B. K. (2016). A case study on partitioning data for classification. *International Journal of Information and Decision Sciences*, 73-91.

Gupta, T. (2021, 12 15). *Deep Learning Feedforward Neural Network*. Retrieved from Towards Datascience: https://towardsdatascience.com/deep-learning-feedforward-neural-network-26a6705dbdc7

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Thirteenth International Conference on Artificial Intellegence and Statistics* (pp. 249 - 256). Proceedings of Machine Learning Research.

Morde, V. (2021, 12 15). *XGBoost Alorith: Long may she reign!* Retrieved from Towards Datascience: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Nabi, J. (2021, 12 15). *Recurrent Neural Networks (RNNs)*. Retrieved from Towards datascience: https://towardsdatascience.com/recurrent-neural-networks-rnns-3f06d7653a85

Kulshrestha, R. (2021, 12 15). *NLP 101: Word2Vec — Skip-gram and CBOW*. Retrieved from Toward Datascience: https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

Brownlee, J. (2021, 12 15). *Hyperparameter Optimization With Random Search and Grid Search*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/

Kumar, A. (2021, 12 15). *Overfitting & Underfitting Concepts & Interview Questions*. Retrieved from Data Analytics: https://vitalflux.com/overfitting-underfitting-concepts-interview-questions/

Yusov, A. (2021, 12 15). *House Prices, first try at predictions: R, xgboost*. Retrieved from Kaggle: https://www.kaggle.com/ayusov/house-prices-first-try-at-predictions-r-xgboost

Jadhav, S. (2021, 12 15). *House Price Prediction Notebook*. Retrieved from Kaggle: https://www.kaggle.com/shraddhajadhav111/house-price-prediction-notebook

Github User. (2021, 12 15). *House price prediction using Xgboost*. Retrieved from Github: https://gist.github.com/gauravgola96/c356acf7b2ae0bdd0673d8e5d303f43e

Castillo, D. (2021, 12 15). *Titanic predictions using keras in R*. Retrieved from Kaggle : https://www.kaggle.com/dylanjcastillo/titanic-predictions-using-keras-in-r/script

Kaggle User. (2021, 12 15). *Predicting house prices using keras with R*. Retrieved from Kaggle: https://www.kaggle.com/floser/predicting-house-prices-using-keras-with-r

Lee, C. M. (2021, 12 15). *Kaggle*. Retrieved from Jigsaw toxic comment classification challenge: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557

Aindow, T. (2021, 12 15). *Deep learning with R: Sentiment Analysis*. Retrieved from Kaggle: https://www.kaggle.com/taindow/deep-learning-with-r-sentiment-analysis

Monkey Learn. (2021, 12 15). *Sentiment Analysis & Machine Learning*. Retrieved from Monkey Learn: https://monkeylearn.com/blog/sentiment-analysis-machine-learning/