Real Estate Price Prediction Using Sentiment Analysis on Real Estate Descriptions

W.E. Molders
STUDENT NUMBER: 2007076

THESIS SUBMITTED IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE &
SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL
INTELLIGENCE SCHOOL OF HUMANITIES AND DIGITAL
SCIENCES
TILBURG UNIVERSITY

Thesis committee:
Prof. Dr. M.M. Louwerse
Dr. D.J. Schad

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2020

Preface

'Obstacles don't have to stop you. If you run into a wall, don't turn around and give up. Figure out how to climb it, go through it, or work around it.' – Michael Jordan

Nothing could be further from the truth, when one says there are no obstacles following the Master Data Science. Running through or jumping over walls like the Hulk, was not in it for me either. It is safe to say that experiencing these obstacles sparks creativity, determination, and a goal-oriented mindset, to solve and leave the obstacles in the rear-view mirror. Although in the background, a basic level of understanding of the different theoretical domains and skills remain a necessity.

I would like to extend my deepest gratitude to my supervisor Professor Max Louwerse, not only for his guidance during the thesis project, but also for sparking my interest in areas of knowledge where they were previously absent. For someone coming from Industrial Engineering and Management, topics and theory about the inner workings of the brain, as well as linguistic features, combined with statistical testing, were relatively new and something that does not share any resembles with previously obtained knowledge. During the premaster's program, the course 'Language, Cognition and Computation' was therefore one of the toughest courses to choose from, but due to the a fair and motivating examination structure and an enthusiastic mentor, it was one of the courses I had to follow. It has earned its spot alongside a one-year class in Biology during high school, in the rank of all the courses that I have been most excited about.

On another note, I would like to thank Bagels and Beans and Kruim, in Gouda, Netherlands, for their support and granting additional stay in there accommodation, as due to the COVID-19 crisis, all public spaces were prohibited for customers for a rather long period of time. Staying alone at home, trying to focus on the task at hand was not always so easy. A large cappuccino with a shot of caramel after the strict hospitality rules were lifted, did miracles. I would also like to thank the community of Github, for providing the hardly needed guidance with small to large coding errors. Especially, after searching for coding mentors was, sadly enough, a process to no avail.

Most importantly, I would like to thank my parents and my girlfriend for their never-ending support. Especially during the last period, where the majority of my time and focus went to the compilation of this research.

Real Estate Price Prediction
Using Sentiment Analysis on Real Estate Descriptions

W. Molders

This study examines the prediction power of sentiments found in real estate description on its corresponding ask price, sold price and price difference. Sentiment is extracted from descriptions by utilizing seven widely used lexicon-based approaches, such as VADER, LIWC15 and SentiStrength. The real estate descriptions and pricings are extracted from the website of real estate agency Zillow, based in the United States. The results showed R^2 evaluation metrics close to zero, for all regression cases. Regression analysis taking all lexicons together scored a higher R^2 evaluation metric compared to models with individual sentiments, but yielded a higher test error on all cases. For classification, Pattern.en produced the highest prediction metric (F^1 = .64) for the price difference class. Amongst all lexicons considered, VADER scores the highest-class prediction power with a mean Macro-F^1 of .36.

1. Introduction

E-commerce has known a tremendous growth and is still growing each year in terms of both revenue and percentage of total global retail revenue. Statistics show a 19.8% and a 16.4% revenue growth in 2018 and 2019 respectively. Due to COVID-19, it is expected that for 2020, the growth percentage will be approximately 10.2%, lower than previously estimated. To provide additional insights into these numbers, the expectations for current year translates into a total global revenue of 2,275,953 million dollars. For 2021, 2022 and 2023 it is prognosed that the positive revenue growth will continue to follow a continuous negative slope, to a 5.0% revenue growth in 2024 (Statista, 2020). Although its general declining trend in total revenue growth, the total share of e-commerce amongst retail sales follows a steady, positive slope from 10.4% in 2017 to an expected 22.0% in 2023 (Oberlo, 2020). Due to this increase in online consumers, it is worthwhile to further investigate the online purchase process.

Research was devoted to numerous aspects which could potentially influence this process. Lee and Shin (2014) concluded that consumer consider their peers' evaluation, such as a product review, when purchasing. It is experienced as a valuable resource to acquire additional information about the products or services they are considering buying. It is also found that the purchase intention of potential buyers is influenced by the quality of the review. High-quality is typically depict as reviews that contain objective information, which are relevant and accurate, whereas subjective information within reviews is believed to be low-quality. Although such findings are suggested, there are no current conclusive findings to re-enforce this statement. Other research suggests that the purchase intention and perceived value among buyers was increased by considering the feedback of other consumers (Huang, Lurie and Mitra, 2009). However, this was mediated by the type of product, whether it related to experience or search goods (i.e. products and services that can be properly evaluated beforehand). This finding is supported by Ghose and Ipeirotis (2011), who found that subjective reviews were found more helpful and

Lee and Shin (2014) also found that recommendations made by reviewers resulted in a higher purchase intention for experience goods when these were subjectively formulated. The researchers also indicate that further research is needed to better understand the underlying mechanics of reviews and their effect on purchase intention. Korfiates, García-Bariocanal and Sánchez-Alonso (2012) make the same suggestion and call for further use of sentiment analysis. Their research found that the perceived helpfulness of reviews is influenced by, in particular, stylistic elements (e.g. similes and metaphors). They also imply that positive and negative reviews each effect the consumer differently. After the emphasis of previous research to conduct further analysis towards the subjective nature of reviews, our research tries to do so by examining the subjective nature of an experience good on its perceived value by both buyer as seller.

One of the branches in which the shift from retail to e-commerce becomes more apparent is in real estate. As early as the year 2000, e-commerce was identified for its potential opportunities and challenges for real estate, emphasizing a new way of communication with potential buyers (Bardhan, Jaffee & Kroll, 2000). Currently, Zillow leads the real estate and rental marketplace with a comprehensive housing database of more than 110 million U.S. homes. The database consists of houses currently for sale and for rent, but also ones that are not currently on the market. Zillow registered a monthly average of unique users of 151.6, 157.2 and 172.6 million in 2017, 2018 and 2019, respectively. The platform connects buyers and sellers with a Premier Agent Partner in order to transition from real estate. Renters are connected with landlord partners and Borrowers are provided with the option to finance with Zillow Home Loans or connect with a mortgage partner. As of 2018, Zillow also launched Zillow Offers, which provides sellers (homeowners) the ability the receive cash offers from Zillow to purchase their home, in order to relieve stress from the seller by eliminating the process that would otherwise follow. After buying the real estate, light, make-ready repairs are performed and placed on the open market (Zillow Group, 2020). By applying such mechanisms, combined with machine learning algorithms and the power of e-commerce, Zillow changes the traditional home selling and buying process (which typically uses a 5-6% home commission). A process that has been previously been resistant to change. Zillow Chief Executive Rich Barton said: 'It's the dawn of e-commerce for real estate' (Dezember & Rudegeair, 2019).

Considering these developments, current study focusses on sentiment analysis within the branche of real estate. Real estate descriptions can be seen as a review, or otherwise a form of advertisement, but nonetheless a valuable resource (Mou, Zhu & Benyoucef, 2020) to obtain additional information of the real estate they are willing to purchase. Since real estate's value can be derived from hedonic indices, one tends to first visit the property before actually making the purchase. Consumers have to validate the estate by inspecting and evaluating it in first person. It can therefore be considered an experience good. Derived from this notion, it is expected that a more subjective real estate description leads to a higher purchase intention and product evaluation (i.e a higher price).

In order to investigate previously stated area of interest, namely whether sentiment in real estate description bares any prediction power towards the real estate's value, we can derive the following research questions:

RQ1: To what degree can sentiment in real estate descriptions predict real estate asking price?

RQ2: To what degree can sentiment in real estate descriptions predict real estate selling price?

RQ3: To what degree can sentiment in real estate descriptions predict price difference in real estate asking and selling price?

Extracting sentiment from text can be conducted in various ways, resulting in different sentiment values given to the same text. Therefore, the study will account for several of the numerous sentiment lexicons that can deployed and compared, such as VADER, LIWC and Bing. Additionally, the predictive power of a combination of the sentiment results, given by the lexicons, will also be explored.

RQ4: Which sentiment analysis lexicon is best able to predict real estate pricing?

RQ5: Does combining sentiment analysis lexicon results yield a greater predictive power?

The structure of the report is as follows: section two focusses on related works with regards to sentiment analysis, sentiment analysis tools and research related to price prediction and real estate. The third section mentions the method used during the research, which includes the description of the dataset, which type of pre-processing methods are applied and which models are used to analyze the data and answer the research questions previously formulated. The results are listed in section four. The discussion of the results will be formulated in the fifth section. The last section mentions the conclusion, as well as limitations and future implications of the study.

2. Related Work

Text not only contains objective information, but also linguistic features that involve subjective meaning such as sentiments, opinions, and attitudes. These are characteristic that together form the basis of Sentiment Analysis. Such an analysis is part of Natural Language Processing, which utilizes different methods to analyze unstructured data. Methods such as tokenization, lemmatization, stop word removal and stemming. The bag-of-words method is used by many analysis tools, including 'Term Frequency – Inverse Document Frequency (TFIDF)'. Latter method improves the bag-of-words method by applying weights. This method has been previously used in research conducted by Stevens (2014), which analyzed identical data used in this research. The research was able to predict pricing indicators above formulated baselines. Our research builds further on this research, by examining the sentiment features found in text by means of an unsupervised, lexicon-based approach.

2.1. Types of Sentiment Analysis

There are different methods used for sentiment analysis. Three distinct approaches can be differentiated in which these methods can be categorized, namely a) Machine Learning based, b) Lexicon based, and c) Hybrid based approaches. Machine Learning based approaches train a classifier applied on manually labelled data. Manually labelling the complete dataset is a rather cumbersome process and the quality of the training data has a high impact on the performance of the classifier. It requires a large database in order for it to be correctly applied an achieve high accuracy. This approach outperforms lexicon-based approaches, which make use a sentiment lexicon aimed at deriving the polarity (neutral, positive and negative) of textual input. The Lexicon based approach is easier to deploy, since there is no more need for a large, manually labelled dataset. The Lexicon based approach can be further divided into Dictionary based approach, which is based on dictionary words (given by WordNet or other entries), and Corpus based approach. Latter uses corpus data and another division can be made between Statistical and Semantic approaches within the Corpus based approach Sadia, Khan and Bashir (2018). In this study, a lexicon-based approach is used in order to conduct the sentiment analysis.

Sentiment classification is generally conducted in two variations, either classes are identified in a binary fashion as 'Positive' and 'Negative', or in a multiclass fashion such as 'Positive', 'Negative', 'Neutral' or a more fine-grained level with more than three classes. Research conducted by Sadia (2016) applied lexicon-based sentiment analysis on restaurant reviews, by means of a unigram language model incorporated with the NRC lexicon, in order to generate a polarity score. Binary class classification achieved a 85.5% accuracy, while multi-class accuracy quickly dropped to a 48% accuracy score, due to increased complexity when applying multi-classes to the model. In order to ensure no quality loss in our data, sentiment polarity scores are not transformed into classes, but rather serve directly as input for the price prediction model.

2.2. Sentiment Analysis history

Sadia, Khan and Bashir (2018) conducted research in order to create an overview of all developments within lexicon-based sentiment analysis. According to the researchers, the year 2001 marks the year in which sentiment analysis in people's opinions is first mentioned, in the research conducted by Das and Chen (2001) and Tong (2001). Although the research of Sadia et al. (2018) does not covers all developments, it does give the general lines of the progress made. Text classification techniques like Naïve Bayes, Support Vector Machines (SVM) were introduced in 2002, word opinion mining was first mentioned in 2003, in 2004 WordNet was used to help distinguish between the semantic orientation of opinion words, clause level sentiment analysis was proposed in 2006, in 2008 a technique was introduced to identify the orientation of context dependent product reviews and in 2011 Semantic Orientation Calculator (SO-CAL) was proposed, which used thesauruses of words and their semantics to assign polarities to text. In 2013, a sentence based opinion lexicon was used for the German language, in 2014 the idea of using a lexicon based approach for multi-lingual sentiment analysis was proposed where input text was translated into reference language, in the same year data pre-processing and information retrieval by means of SVM were described and in 2015 sentiment analysis on twitter was conducted by using SentiWordNet and utilizing the Hashtag Sentiment Lexicon and the Sentiment140 Lexicon (Rosenthal, Nakov, Kiritchenko, Mohammad, Ritter & Stoyanov, 2015). 2014 In 2016, the classifiers Vader, SentiStrength and SentiWordnet were utilized in a approach for new meta-level features for sentiment analysis. Meta-level features are usually manually designed and derived from other features, detect intrinsic relationships among pairs (document, class) or a triple (document, class, algorithm) (Canuto, Gonçalves, & Benevenuto, 2016). Sadia (2016) build on top of latter research and applied the National Research Council Canada (NRC) Word-Sentiment Association Lexicon. In 2017, the WKWSCI Sentiment Lexicon was proposed by Khoo and Johnknan (2018), during which five other prevailing lexicons were compared: Multi-Perspective Question Answering (MPQA), NRC, Hu & Liu Opinion Lexicon, SO-CAL, Subjectivity Lexicon and General Inquirer and Word Sentiment Association Lexicon (GIWC). WKWSCI, MPQA, Hu & Liu and SO-CAL are equally appropriate for product review sentiment categorization, which obtained an accuracy rate between 75% and 77%. The researchers recommend the Hu & Liu lexicon for product review texts and the WKSCI for non-review texts**.** SentiWordNet was also tested, but obtained the worst accuracy of 65%, GIWC and NRC scored similar scores. All Lexicons were processed with and without a simple method handling negation, if a negation word occurred preceding a sentiment-bearing word (with up to one word in between), the polarity of the sentiment is reversed. Noticeable, is that applying such a simple negation method, boosted performance by up to 5%. Although negation handling showed an improved score, the ratio between performances of the Lexicons appear to remain equal.

Due to such scores mentioned above, NRC will be used as baseline value during this study. Since a Sellers description can also be seen as a review and de Hu & Liu Opinion Lexicon is well known for its application in regard to such domain, it will be taken into our sentiment analysis methods application. Out-of-the-box methods such as above will be applied, without adding any additional negation handling algorithms.

2.3. Sentiment Challenges

Sentiment analysis knows several challenges, as it is a complex task to perform. One of which is identifying whether a word is subjective or objective. Sadia et al. (2018) provides the following example: A. 'The customer's language was very crude', where the word 'crude' can be identified as an opinion word, B. 'Crude oil is being imported', where 'crude' can be identified as objective. Another challenge within sentiment analysis is its domain dependence, the same word can have different semantic meanings across different fields. Shaukat, Hameed and Luo (2020) conducted research towards domain specific lexicon generation through sentiment analysis, in which their goal was to identify words that are interpreted differently in various fields (Politics, Terrorism, Life, Science, Sports and Movies). They found that the domain specific word 'Growth' was labelled positively in Politics and Life, but negatively in Terrorism and Science. Another example includes the word 'Pay', which is negatively labelled in Politics, Sports and Movies, but positively in Life. Interestingly, it was also found that the distribution of Negative/Positive words within the different datasets were not equal, e.g. the Politics, Gossips and Movies datasets gave more than twice as much positive related words as negative, while for Science the opposite was true. Yet another challenge is detecting sarcasm and contextual meaning in text, since sarcasm utilizes positive words, but in a negative context. Each method applied in this research utilizes their own measurements to tackle these challenges, giving different polarity scores as a result. In order to capture sentiment correctly, by means of different lexical methods, we will take several into consideration when validating sentiments prediction power on real estate pricing.

2.4. Sentiment Lexicons

There are a wide range of different Sentiment Lexicons which can be used to extract and determine the sentiment in the different descriptions. Extended evaluation has been conducted by the researchers Ribeiro, Araújo, Gonçalves, Gonçalves and Benevenuto in 2016. During this evaluation, 24 sentiment analysis methods were benchmarked based of 18 labelled datasets, which covered text posted on social networks, movie reviews, product reviews, and in news articles. The majority of latter mentioned researchers were also involved in previously conducted sentiment analysis benchmarking studies (Gonçalves, Araújo, Benevenuto & Cha, 2013; Araújo, Gonçalves, Cha & Benevenuto, 2014). Their latest study is more comprehensive and more up to date (e.g. both LIWC2007 and LIWC2015 are considered) and will therefore be mainly consulted. Another extensive benchmarking study was performed by Abbasi and Dhar (2014), where twenty tools (stand-alone and workbench tools) were tested on five different Twitter datasets (Pharma, Retail, Security, Tech and Telco). However, Ribeiro et al. (2016) use more distinct contexts and solely unsupervised methods (comparing both would be unfair due to extensive parameter tuning and validation during training).

Lexical methods vary largely to the context from which they were derived. Although this is the case, researchers tend to accept any popular sentiment analysis methodology, without exactly knowing the relative performance of the different methods and their advantages, disadvantages, and limitations. Even more so across different contexts. During the comparison study of Ribeiro et al. (2016), off-the-shelf methods are considered,

which excludes supervised methods that require labelled sets for training. Latter has the ability to adapt and formulate models that are specifically trained for certain purposes and contexts**.** Our research will follow similar structure, supervised approaches are outside the scope of this study.

During the comparison made by Ribeiro et al. (2016), an equal playing field was created to evaluate the word lexicons specifically. Sentence-level analysis methods such as VADER and SO-CAL use intensifiers, punctation transformation, emoticons, and other heuristics in order to award sentiment polarity to a sentence. The same techniques used for VADER are applied when using other Lexicons, which do not apply the same rules in their original state. VADER is widely used and is even applied in the well-known NLTK python library, their heuristics based on grammatical and syntactical cues go beyond the bag-of-words model. Concretely, rules considering punctuation, capitalization, degree modifiers, constructive conjunction and tri-gram examination for negation are crossed over to the other methods. Results were drastically improved for most of the lexicons for sentence-level sentiment analysis (Ribeiro et al., 2016). However, during this research, the cross-over of heuristic features will not be conducted. Lexicons will be evaluated as is, to maintain their own characteristics when used out-of-the-box.

The main findings of the study conclude that there is no single method that achieves the best performance over all the different datasets. Moreover, prediction performances vary considerably across datasets and sentiment analysis on the same data yields different results by the different methods. The researchers stress that it is crucial to test different methods in a sample set of the data, before applying a method that is accepted by the research community. Another key takeaway is that, although several methods score high on accuracy and Macro-F^1 on certain datasets, the coverage of many of those methods are low, such as those of VADER and SentiStrength. Taken together, the researchers suggest that practitioners need to take the trade-off between prediction performance and coverage into account. Another aspect worth noting, is that LIWC was not originally developed to detect neutral sentences, but that the updated version of LIWC scores amongst the top five methods for both two-class (positive and negative) and three-class (positive, negative and neutral) sentiment prediction. In general, the benchmarking test found that most methods are more suited to classify positive than negative or neutral sentences, and also suggests that some methods may be more biased towards positivity. Additionally, neutral words showed to be even harder to be detected by most methods. Since we will not transform the sentiment polarity scores into classes, such knowledge does not have to be taken further into account, but it still worth noting while doing any sentiment analysis.

Based on Macro-F^1, for the two class experiments, the top nine methods consisted out of SentiStrength, Sentiment140, Semantria, OpinionLexicon, LIWC2015, SO-CAL, AFINN, VADER and Umigon. All of which, except SentiStrength, are also in the top nine of the three-class experiment. Pattern.en replaces the SentiStrength method. In situations, where a preliminary evaluation is not possible or not done, these methods would be preferable. Interestingly, VADER did not end first position in any of the datasets but was most consistent overall.

Since the different methods all yield a different sentiment score, it could be of interest to further investigate whether different sentiment analysis tools are better in predicting real estate pricing as opposed to others. After previous findings, the methods that will be used during this research are SentiStrength, Pattern.en, OpinionLexicon, LIWC2015, AFINN and VADER. In accordance, Abbasi and Dhar (2014) also found that SentiStrength yielded the best overall accuracy on average over the five different test sets. For baseline purposes, we will also utilize EmoLex (NRC Emotion Lexicon), which ended mid-tier in the benchmark study (position 14 in two-class and position 10 in three-class). In order to build on these findings and provide a proper starting point for potential, follow-up

research, extended models such as SentiStrength-SE (domain specific SentiStrength for Software Engineering)(Islam & Zibran, 2018) or LIWBC (Linguistic Inquiry and Word Bigram Count) (Carvalho, Rodrigues & Guedes, 2018) are not taken into consideration.

2.4.1. LIWC

Linguistic Inquiry and Word Count (LIWC) is a dictionary-based computational tool, which was first developed in 1993 as part of an exploratory study of language and disclosure. The second and third editions extended the dictionary (Pennebaker, Francis, & Booth, 2001; Pennebaker, Chung, Ireland, Gon- zales, & Booth, 2007). LIWC has been used as benchmark in numerous studies in which a new lexicon was proposed, such as in the research for the development of VADER and has been used extensively in the social media domain. The LIWC Lexicon has been previously favoured by professionals due to its extensive validation and accessible and straightforward implementation (Gilbert & Hutto, 2014). In 2015, the dictionary was again extended, now outputting approximately 90 variables, under which: summary language variables (e.g. 'emotional tone'), general descriptor categories (e.g. 'words per sentences'), standard linguistic dimensions (e.g. 'percentage of words that are pronouns in the text'), word categories tapping psychological constructs (e.g. 'cognition'), personal concern categories (e.g. 'leisure activities'), informal language markers (e.g. 'fillers', such as 'uh', 'you know'), and punctation categories. During our research, these variables do shed additional insights into the text, but the main concern is sentiment extraction and can be further disregarded. The default LIWC2015 dictionary consists of around 6,400 words, word stems and a select number of emoticons (Pennebaker, Boyd, Jordan & Blackburn, 2015). Latter part, emoticons, has been added in the 2015 edition. These are basic punctuation-based emoticons, such as ':)' and ';('. The dictionary also added words that are frequently used in social media and text messaging (e.g. 'bae', 'lol') under the 'Netspeak' category, something that was missing opposed to other lexicons. Additionally, by means of more current statistical method, words were omitted that were not classified appropriately and words were added that were previously missing (Pennebaker, Booth, Boyd & Francis, 2015). Word categories 'Regular Verbs', 'Sexuality' and 'Fillers' and 'Assent' in Informal Speech had the greatest change in opposed to the 2007 version. LIWC awards a sentiment score by incrementing the scale scores of the different sub dictionaries in which the word is present. An example given is 'cried', which is retrieved from five-word categories: 'sadness', 'negative emotion', 'overall effect', 'verbs', and 'past focus' (Pennebaker, Boyd, Jordan & Blackburn, 2015). LIWC text-analysis is computed by counting words in defined dimensions, one of the advantages is that users can develop their own dictionaries which can be applied to these dimensions. In such manner, any aspect of language use can therefore be analysed, for example to study the Linguistic Category Model (LCM) (Seih, Beier & Pennebaker, 2017). LIWC has also been implemented in numerous languages, such as Dutch (Boot, Zijlstra & Geenen, 2017; Saleem, 2017; Tulkens, Hilte, Lodewyckx, Verhoeven & Daelemans, 2016), German (Timasjan, Sandner & Welpe, 2010), Spanish (Salas-Zárate, López-López, Valencia-García, Aussenac-Gilles, Almela & Alor-Hernández, 2014), French (Piolat, Booth, Chung, Davids & Pennebaker, 2011) and Chinese (Zhao, Jiao, Bai & Zhu, 2016). By taking this lexicon into consideration, findings of this research has the potential to be easily applied in different formats (language, domain, etc.)

2.4.2. VADER

VADER is a lexicon and rule-based sentiment analysis tool that is attuned to sentiments in social media, compiled by the researchers Gilbert and Hutto (2014). As previously stated, LIWC has been used extensively in social media. However, it was not specifically meant to

do so. VADER retains and extends the benefits of the traditional lexicon LIWC. It differentiates itself from LIWC due to increased sensitivity in sentimental expressions in social media contexts, whilst also be superior in generalization to other domains. As noted, LIWC2007 has been improved by means of the LIWC2015 edition and added features that could now also capture parts of sentiment in social media that was previously unable to (such as incorporating basic emoticons). Benchmarks in Gilbert and Hutto (2014) are therefore no longer representative for the LIWC Lexicon. The source code to determine the polarity scores reveals that lowercasing, punctuation, negation, emoticons and other processing techniques are considered (NLTK, 2020; Hutto, 2020). One of the other techniques include the detection of capitalized words, which increases polarity with an empirically derived mean sentiment intensity rating. Equivalently, a positive intensity rating is applied for booster words (e.g, 'absolutely', 'amazingly') and negative intensity rating for dampening words (e.g. 'kind of', 'sort of'). Laden idoms (e.g. 'break a leg') and special cases (e.g. 'the shit', 'the bomb') are also considered. VADER outputs positive, negative and neutral Word Sentiment variables. Additionally, a compound rate is calculated based on these variables, which will be used in our analysis. VADER uses over 7,500 lexical features in order to formulate the values for the latter mentioned outputs. Guerrero, Olivas, Romero and Herrera-Viedma (2015) also applied the VADER Sentiment Analysis and mention 'okay' (+0.9), 'good' (+1.9), and 'great' (+3.1) as examples of positive valance words, likewise, 'horrible' (-2.5) is considered a word with a negative valence, ':(' is considered a negative frowning emoticon and gets the value (-2.2), slang is also taken account in for example 'sucks' or 'sux' (-1,5). The compound rate is calculated according to a set of rules and normalized between -1 and 1. Although the focus of VADER lied on evaluating sentiment of tweets, Gilbert and Hutto (2014) were able to establish a parsimonious rule-based model that outperforms individual human raters and generalizes across context more favourably than other benchmarks in their research (including LIWC, ANEW, the General Inquirer, SentiWordNet and numerous machine learning oriented techniques). It is therefore of benefit to measure its sentiment analysis polarity score in relation to price prediction, although it being in another domain other than the context it was evaluated in (Social Media Text, Amazon Reviews, Movie Reviews and NY Times).

2.4.3. Pattern

The Pattern Lexicon was first introduced by Schmedt and Daelemans (2011), in name of CLiPS (Computational Lignuistics and Psychologuistics). The Pattern library offers a complete NLP framework, with functionality for web mining, natural language processing capabilities (tagger/chunker, n-gram search, sentiment analysis), machine learning and network analysis. The package is organized in such a way, that the modules can be chained together, e,g. crawling text from Twitter (Pattern.web), followed by parsing via the part-of-speech tags (Pattern.en), querying by syntax and semantics (Pattern.search), and eventually training one of the possible machine learning classifiers (Pattern.vector). Our research restricts its utilization to the Pattern.en module, which encompasses a fast, regular expressions-based shallow parser for the English language. It is able to identify sentence constituents (nouns, verbs, etc.) and uses a finite state part-speech tagger, combined with a tokenizer, lemmatizer and chunker. The Pattern Lexicon can also be applied for the Dutch language, since it also contains the module pattern.nl, which uses the BRILL-NL language model (Smedt & Daelemans, 2012). The official Github repository of Pattern contains the subjectivity-based lexicon database (Schmedt, 2014), which reveals metadata information of each word used in the Lexicon, such as WordNet corpus identifiers, part-of-speech tags, word sense and polarity and subjectivity scores. An

example line from the database is ''<word form="abundant" cornetto_synset_id="n_a-525828" wordnet_id="a-00013887" pos="JJ" sense="present in great quantity" polarity="0.6" subjectivity="0.9" intensity="1.0" confidence="0.8" />". The Pattern Lexicon can be utilized using the framework Textblob (Sarkar, 2019) which uses identical implementation (Loria, 2017), but also has its native Pattern Python library, which will be used in this research TextBlob is built on top of NLTK and the Python library. An example study which uses the same manner of utilization of the Pattern Lexicon (by means of the Pattern Analyzer function) is in sentiment analysis of twitter user data on legislative assembly election (Singh, Gupta, Singh, 2017).

2.4.4. SentiStrength

SentiStrength was introduced in research conducted by Thelwall, Buckley, Paltoglou, Cai & Kappas (2010), in which its main focus was to detect sentiment in short informal text. It is a lexicon-based classifier with added rules and linguistical features and outputs polarity two values, for positive and negative strength with values ranging from 1 (being no sentiment) to 5 (being strong sentiment). The SentiStrength library gives several options as to its output, namely 'scale', 'dual', 'binary', 'trinary'. 'Scale' is most consistent with other methods and will therefore be preferred over the other options. Original key features consist of a sentiment word list including human polarity and strength, of which some stemmed, a spelling correction algorithm which deletes letters that are repeated more than proper English dictates, a booster word list, an idiom list, a negation word list, a strength boost when at least two letters are added to a word (such as ''haaaappy''), an emoticon list with polarities, a minimum sentence positive strength of two when there are exclamation marks (unless negative), a sentence positive strength boost of one when there is repeated punctation with one or more exclamation marks and negative sentiment is ignored in sentences. These features have been improved with SentiStrength 2, which main focus was to address the relative weak performance of the method to detect negative sentiment strength (Thelwall, Buckley & Paltoglou, 2011). The main take-aways are that the sentiment word list was extended with negative General Inquirer (another sentiment lexicon) terms along with their human-coded sentiment weights and stemming, the sentiment word terms were checked for incorrectly matching words and derivative words that did not correspond, which resulted in terms being converted to wildcards (multiple variants) and that negating negative terms do not make them positive but neutral (e.g. 'I do not hate him', is processed as neutral). In addition, an extension to the idiom list was made, where phrased were added that indicate word senses for common sentiments words. An example given is 'is like', where 'like' is not positive but used as comparison word, so must be awarded a neutral sentiment polarity. Such a manner is a relatively simple alternative to award sentiment to word contexts, in contrast to using POS-tagging technique. The lexicon consists of around 2,500 terms. SentiStrength has also been applied in different language settings, such as Turkish (Vural, Cambazoglu, Senkul & Tokgoz, 2013) and Spanish (Vilares, Thelwall & Alonso, 2015; Baviera, Sampietro & García-Ull, 2019). It is also incorporated in several software suites, such as Mozdeh, COSMOS and Chorus (SentiStrength, n.d.). More recent studies in which the lexicon has been applied is in examining gender bias in sentiment analysis (Thelwall, 2018), in aggressive language identification (Orasan, 2018) and in political conversations (Baviera et al., 2019).

2.4.5. Opinion Lexicon (Hu & Liu)

The 'Opinion Lexicon' is also referred to as ''Bing Liu's Lexicon', 'Hu and Liu's Lexicon' and in the used R package as 'Bing'. The lexicon is divided into two files, one with the

positive words and one with the negative words. It contains a total of around 6,800 words, of which approximately 2,000 words or phrases that are positive and 4,800 words or phrases negative. These lists have been compiled over many years and started with the first research conducted by Hu and Liu (2004). The Opinion Lexicon also utilizes a list of comparative words, which is also compiled over the years and started with the first paper of Jindal and Liu (2006). This list consist of non-standard words that indicate comparisons (e.g. 'beat, defeat', 'equally'), more specifically nonequal gradable comparisons (e.g. 'ahead of', 'blow away', 'compare X and Y') and equative comparisons (e.g. 'as.. as', 'the same as'). Words are awarded a score between -1 (negative) to 1 (positive) (Nadil, 2019). One of Liu's recent studies incorporates the Opinion Lexicon for research in lifelong learning memory networks (Wang, Mazumder, Fei & Liu, 2018).

2.4.6. NRC Emotion Lexicon (EmoLex)

The NRC Word-Emotion Association Lexicon (EmoLex) assigns a sentiment type to words, as well as sentiment It considers the categories anger, anticipation, disgust, fear, joy, sadness, surprise and trust. The lexicon consists of approximately 14,000 words and covers approximately 25,000 word senses (Mohammad & Turney, 2010; Mohammaed & Turney, 2013).The association scores are binary (associated or not). For each word in the sentence, the specified word will be matched with the categories and for each occurrence, the category will be awarded a +1. Instead of positive and negative words with their corresponding algebraic score, each sentence gets a score for each category.
NRC offers different lexicons, manually created lexicons include 1) NRC Word-Emotion Association Lexicon (or NRC Emotion Lexicon/EmoLex), 2) NRC Valence, Arousal, Dominance Lexicon, 3) NRC Affect Intensity Lexicon and 4) NRC Word-Colour Association Lexicon. There are also automatically generated lexicons, which include 1) NRC Hashtag Emotion Lexicon, 2) NRC Hashtag Sentiment Lexicon, 3) NRC Hashtag Affirmative Context Sentiment Lexicon and NRC Hashtag Negated Context Sentiment Lexicon, 4) NRC Emoticon Lexicon (Sentiment140 Lexicon) and 5) NRC Emoticon Affirmative Context Lexicon and 6) NRC Emoticon Negated Context Lexicon (Mohammad, 2018). During the benchmarking study of Ribeiro et al. (2016), EmoLex, NRC Hashtag and Sentiment140 were tested. EmoLex ended mid-tier during the benchmarking study and lower tier overall compared to the other chosen lexicons in this paper. EmoLex will be utilized via the Syuzhet package (Jockers, 2019).
EmoLex is available for over a hundred languages since 2017, enabled by Google Translate. It has been implemented in emotion analysis, abusive language detection and personality trait identification, amongst others. More recently, NRC has been used for emotion analysis in tweets during the Covid-19, a creative poetry language generation system and speech analysis of the Mexican President and other politicians (Mohammad, 2020).

2.4.7. AFINN

Finn Arup Nielsen created the AFINN lexicon as the AFINN Word Database (Nielsen, 2011). The AFINN lexicon includes internet slang and obscene words. A set of obscene words formed the starting point of the database and was continuously expanded by researching posts on Twitter and sets of words from the Urban Dictionary and Wiktionary to account for acronyms and abbreviations. The lexicon awards a score range between -5 (negative) to 5 (positive), with eleven values in between (Nadil, 2019). The lexicon also considers emoticons (Nielsen, 2017). It was first considered to apply the AFINN lexicon

via the Syuzhet R package. However, the package utilizes an older version of AFINN, namely AFINN-en-111. AFINN-en-96, the oldest version, encompasses approximately 1500 words and phrases with their associated polarity score, whereas version 111 has an increased lexicon of around 2500 words and phrases compared to its predecessor. The latest version, AFINN-en-165, has an even bigger lexicon with over 3,300 words and phrases. A clear limitation of AFINN is that there are no multiple coders that rate the same words, as well as that the 3,300 are not unique words but are also inflections of the same word. AFINN is applied via the bag-of-words principle, where words are treated as independent entities and context is disregarded. Frequency of words are still taken into consideration. Some pre-processing techniques have to be applied, such as tokenizing and transforming text to lowercase. In opposed to some other dictionaries, where stemming is needed in the normalization process, AFINN does not need stemming as multiple inflections of the same word are incorporated into the dictionary (Enevoldsen & Hansen, 2017). Example studies of where the AFINN lexicon was used includes sentiment classification in stock prices (Urolagin, 2017) and sentiment analysis in restaurant rating (Gan & Yu, 2015). The author created a wrapper library in Python called afinn, which will be used to utilize this method in our research.

2.5. Price Prediction

Numerous studies have examined the relationship between house pricing and its characteristics (Sirmans & Macpherson, 2005; Peterson & Flanagan, 2009). Fundamentally consisting of the hedonic pricing model, which takes general house specifications into consideration, such as the number of beds and bathrooms, but also the size of the house or its lot. Housing can be seen as a heterogenous object, since most characteristics differentiate from one house to the other. Besides the house structure itself, location and neighbourhood are also variables to take into consideration (Ansah, 2011). Our research will account for location, since description has to be made as homogenous as possible. Contents of descriptions are related to its characteristics, our goal is to exclude sentiment that is being captured due to characteristics that discriminates highly from other real estates (i.e. ensure descriptions contents are solely addressed to house features equal to others).

Ling, Ooi and Le (2014) examined the influence of nonfundamental-based sentiment in price dynamics, directly linking the dynamic relation amongst market wide sentiment, the change in housing prices and the market liquidity. Their research shows strong evidence that house prices are affected by sentiment amongst different market participant (Buyers, Sellers and Financial Institutions). In our research, emphasis is laid on the Sellers sentiment manifested in product description. This research also indirectly shows that house pricing is affected by time, since market conditions are continuously changing. Therefore, time constraints are opposed to the dataset (further described in method section).

3. Methodology

During this section, the used dataset will be described, which cleaning, and feature transformations took place and how the different sentiment analysis tools were utilized. Feature extraction was needed not only for sentiments, but also for real estate pricing since they were not correctly gathered. Sentiment features are extracted on sentence level, while analysis is considered by their cumulative sentence polarities (i.e. sentences are analysed and aggregated towards a document level score).

3.1. Data

The data that will be analyzed during this research is compiled by Nanne van Noord, which scraped the official website of Zillow.com in November 2012. The data comes in a .txt file and are comma separated. The data consists of merged scraped batches and contains 620,143 of rows in total, each indicating specific real estate located in the US. These listings were not necessarily listed for sale, nor recently sold, as revealed by their recorded history. NA values are listed differently amongst the variables, marked with either "–" or "". Data rows are mostly not complete, lacking characteristics that would fully describe the respective real estate. This is most likely due to the fact that most listings are not directly on Zillow (12,727 in total that are listed). A full description of all the variables in the dataset, provided with an example, can be found in Appendix A.

3.2. Pre-processing techniques

Angiani, Ferrari, Fornacciari, Lotti, Magliani and Manicardi (2016) conducted research were sentiment classification was performed on two different datasets, in which different situations of pre-processing were evaluated. These situations included pre-processing 1) basic (i.e. only punctuations), 2) Stemming, 3) Stopwords, 4) Negation, 5) Emoticon, 6) Dictionary and All preprocessing. It was found, that solely applying stemming yielded the highest prediction power. In this research, we evaluate the different sentiment analysis methods out-of-the-box. This means that there are no pre-processing techniques applied before sentiment analysis, except basic cleaning (such as '...More\xa0Less' in text, which was redundant and other misplaced symbols due to scraping and conversion).

3.3. Data Cleaning

First, all rows including 'Description from Zillow' were excluded. This, because these are auto generated by Zillow themselves. Followed by the exclusion of data before 2011, since we partially want to have a homogeneous (i.e. constant) influence of market sentiment (Loi et al., 2014). We denote a timeframe of a maximum of two years as most appropriate, since a smaller timeframe would yield even less rows to conduct analysis on. Text smaller than 50 characters were excluded, since these were rather to small to convey any practical sentiment. Moreover, after sentiment analysis, LIWC15 awarded insanely high values (around 100) for such small text (e.g. 'Beautiful!!!' and 'AWESOME'). The reason as to why, is for us not known. Other sentiment outliers that were excluded also included high sentiment scores, which were generated due to wrong punctuation. Sentences were not constructed as commas were used in the whole document and was therefore seen as one sentence. Rows from the bottom region (below $20,000) of pricing, asking and selling price are excluded from the dataset. After inspection, the dataset contains rent options, which are outside the scope of this research. Listings with a sold price of $1.00, extracted from the History data during feature creation, was also removed. Exclusive housing (e.g. sold price of $109,000,000 and ask price $2,000,000,000) are excluded from the dataset, since they fall outside the range of the z-score (3). Corresponding descriptions were manually inspected and did not find any extraordinary description sentiment.

3.4. Feature Extraction

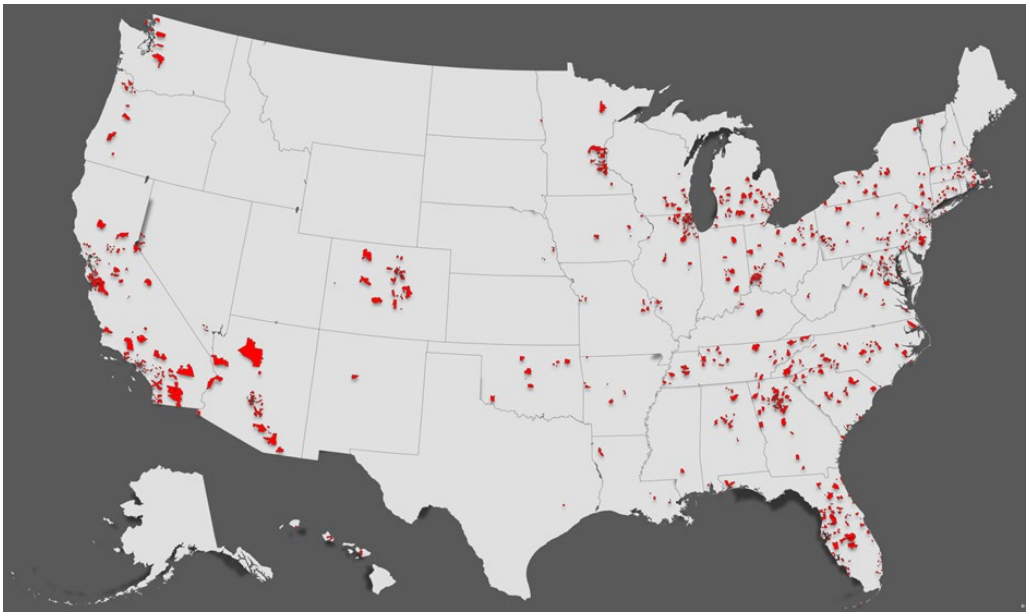
Table 1: Features created.

Ask price	$199,999.00
Ask price class	'Medium',
Ask price date	'06/07/2012'
Sold price	$190,000.00
Sold price class	'Medium',
Sold price date	'09/10/2012'
Price difference	-4.999%
Price difference class	'Decrease',
Postal code	8052,
City	'shade maple',
State	'NJ',
AFINN	1.4000
BING	1.0000
NRC	1.2000
VADER	0.9935
Pattern.en	0.4441
SentiStrength	1.2000
LIWC15	0.6027

Several features needed to be extracted from the original raw data (see table 1), some of which, by additional conditions. Sold price (depicted as 'Price' in the original data) was extracted out of the History information (were all data entries are logged), under the condition that 1) ask price is also mentioned, with a corresponding time stamp before the timestamp of sold price and 2) the sold price timestamp is not prior to 2010. In the original data, pricing was included from 1982 for example. The sold price data needs to be later than the ask price data, since we want to measure whether the description (formulated by the Seller, at a given moment) also predicts the price difference. New features were created, by categorizing ask price and sold price into three classes, 'Low' (price < 150,000), 'Medium' (150,000 < price < 300,000) and 'High' (price > 300,000). Price difference was created by dividing ask price by sold price and stored as percentages. The variable was categorized into 'Decrease' (percentage < 0%), 'Equal' (percentage = 0) and 'Increase' (percentage > 0%). When generating the sentiments, LIWC15 yielded a total document polarity score. Since other methods retrieved sentence level valuations, this polarity score was divided by the division of the number of sentences and the words per sentence of that document. No further normalization amongst predictors was applied.

Since house pricing is related to location and neighborhood, we want to create a more homogeneous dataset by filtering on these conditions. Postal code was extracted from the City State documentation (last 5 numerical values), which can later be used to create subsets to train on. Subsets smaller than 100 were discarded, to ensure a proper sample size. Figure 1 shows the spread of the data considered based on location (size does not present sample size).

Figure 1
Postal codes (with sample size > 100).



The dataset holds 369,536 rows, before postal code filtering. For both asking price and sold price approximately 300,000 rows were used for analysis, after also dropping NA values in their separate dataset. 1200 iterations, or unique postal codes, were performed during modelling.

The classes for ask and sold price, 'Low' (respective class size of 38.06% and 41.25%), 'Medium' (respectively 32.52% and 31.65%) and 'High' pricing (respectively 29.43%, and 27.09%) can be considered balanced (see figure 2). Price change was categorized in 'Increase', 'Equal' and 'Decrease'. However, exploratory analysis shows imbalanced classes for Increase (16.09%), Equal (6.65%) and Decrease (77.26%), see figure 3.

Figure 2
Price class sizes for ask price and sold price.

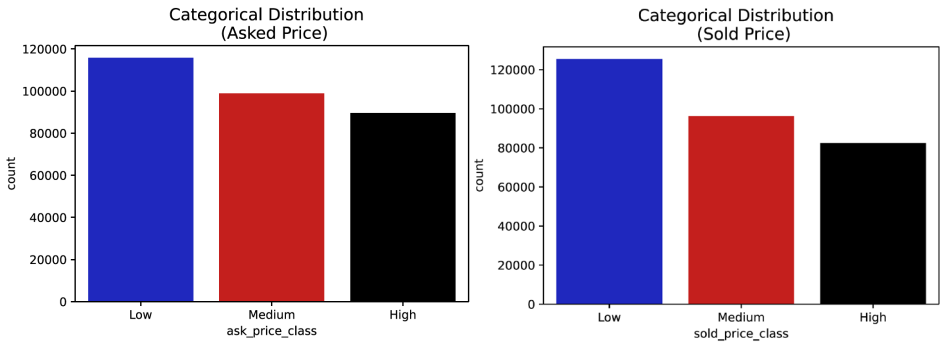
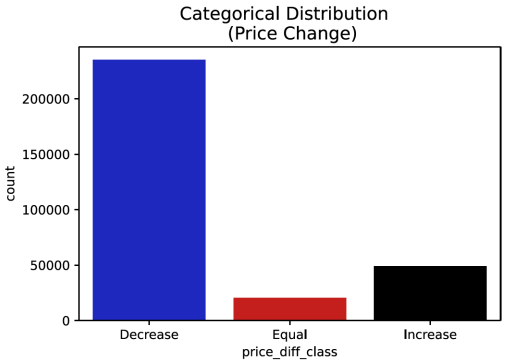


Figure 3
Price class sizes for price difference.



Since equal class only represents 6.65% of the total dataset, methods to eliminate the imbalanced classes problem would mean that more emphasis is placed on only those number of rows, which would not be representative for the population. Moreover, these rows would be consistent with their postal codes, which would limit our research to only those codes. Latter should be avoided since Ansah (2011) mentioned that house pricing is related to their location. Therefore, the class 'Equal' will be discarded and the classification now only considers 'Increase' and 'Decrease'. However, the two remaining classes remain unbalanced. There are several methods to counter imbalanced classes, such as down- or up-sampling and penalizing. Up-sampling is applied for the 'Increase' class. After doing so, the number of rows increased to 469.950 to match classes. After which, postal codes are filtered out which have less than 100 rows, reducing the dataset back to 437.774 rows, with 1197 unique postal codes and an average size of 365 rows. Postal code was also performed for the classification data of the other two variables, which yielded 270.007 rows to be further analyzed on, with 2802 unique postal codes with an average sample size of 109 rows.

Imbalanced data should also be considered within each subset. By doing so, we discard postal codes where each class had less than 20 rows in the subset for ask price class and sold price class. By doing so, the dataset suitable for asking price shrank to 103.989 rows and for sold price 104.694 rows. Although a considerable reduction, the dataset size is deemed appropriate, and it is chosen to not further apply up- or down-scaling within each subset. By doing so, analysis is performed on actual data, but rather on a less large dataset. After latter processing, the classifiers for ask and sold price had 370 postal code iterations, with each having an average size of 281 rows.

3.5. Sentiment Analysis Tools

Pattern, SentiStrength and AFINN are utilized via their native Python library and VADER via the Python library NLTK. The R sentiment package Syuzhet is used to utilize OpinionLexicon and NRC (EmoLex). The package is also able to handle AFINN but does not use the most current version. The default of the package is the Syuzhet Lexicon, which has not been mentioned in previous benchmarking studies. It was developed in the Nebraska Literary Lab and it consists of approximately 11,000 words, accompanied with their corresponding sentiment polarity (Jonkers, 2017). The Lexicon has more negative words than positive, approximately 7,200 negative and 3,600 positives. The method should be better tuned to fiction, since it is taken from a small corpus of novels. Since there is no further indication that this method would yield better results than the already listed methods, the Syuzhet lexicon is not included in the analysis.

3.6. Exploratory Analysis

Other interesting exploratory analysis can be seen in the graphs below, where the features correlations with the target values are plotted. Although not our main focus, the first graph (figure 4) shows the highest correlations amongst selected variables for ask and sold price, which are the number of beds and the number of baths. Most interestingly to see, is the correspondence of the sentiment polarity scores. It provides a brief overview of sentiment analysis tools which, on output bases, relate to each other (figure 5). For example, AFINN and Bing share the most positive correlation, and show resembles to the NRC lexicon. On the other hand, LIWC, VADER and Pattern are more alike when capturing sentiment. Perhaps most noticeably, SentiStrength shows the most correlation between all of the other lexicons (except for NRC). This could well be a hint as to why SentiStrength performed well benchmarking studies (Ribeiro et al., 2016)

Figure 4
Correlations between independent hedonic variables and target variables.

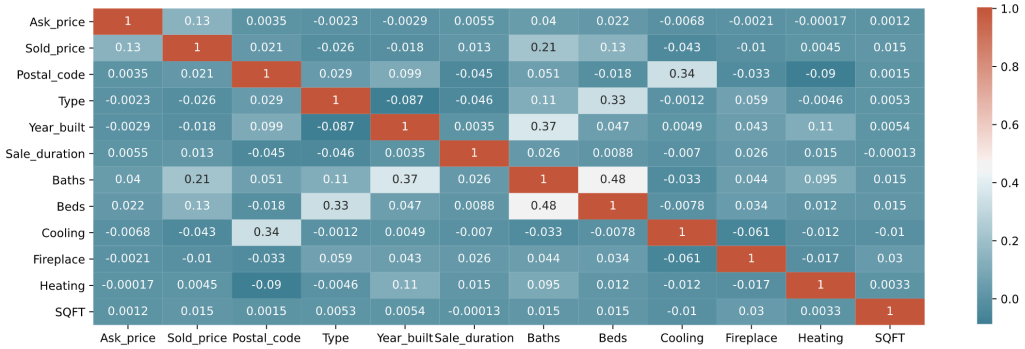
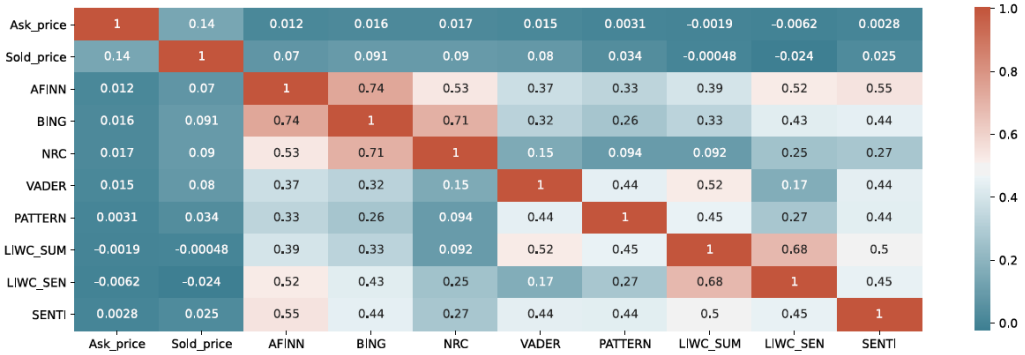


Figure 5
Correlations between sentiment polarity scores and target variables.



3.7. Model parameters, training and testing

For classification, training and testing is done with k-fold cross validation. With k = 5, utilized by a SVM(.SVC) model with gamma set to 0.01 and kernel 'rbf'. An additional classification model was run, KNeighbors with n set to 7, leaf size to 30, power parameter for the Minkowsi metric set to Manhattan distance and the algorithm set to 'kd_tree'. The classification models are utilized via the Python library scikit-learn, where also our model selection was derived from ('Choosing the right estimator', n.d.).

For regression, the train-test split is applied with 20% percent test size. A simple linear regression model with default settings was utilized via the scikit-learn library, along with the OLS linear regression of the Python library package Statsmodels. Latter function enables to retrieve additional statistics of the linear model fit, such as β and p value. A multiple linear regression was used to assess the combination of all the sentiment analysis results on the three different dependent variables.

3.8. Evaluation

Regression evaluation was performed by registering the R^2 values for each individual postal code subset session, which were afterwards averaged. R^2 measures how well future samples are likely to be predicted by the model. Values for R^2 can range from negative values to 1.0, where 1.0 indicates that there is no error in the regression analysis. A zero value means that the regression model predicts mean values of the target values, whereas negative values means that the model produces more errors than simply taking the average. For the prediction evaluation of the model, RMSE is used as evaluation metric.

For classification, the Macro-F^1 values were registered (averaging over F^1 values for class 'Increase' and 'Decrease') for each individual run for postal code. Afterwards the average was taken. A score of 1 means perfect precision and recall, 0 means lowest precision and lowest recall.

4. Results

Table 2
Regression model performance with β, p value, R^2 and RMSE for predicting real estate asking price. β, p value, R^2 were derived from the training data, RMSE was calculated based on the test data. Values are sorted by RMSE value, in descending order.

Feature	β	p value	R^2	RMSE
BING	41,781	**0.203**	**0.029**	**191,956**
NRC	37,518	0.204	0.028	192,303
VADER	56,868	0.232	0.023	192,644
AFINN	15,632	0.275	0.021	192,766
LIWC15	-12,066	0.335	0.012	193,365
SentiStrength	8,593	0.380	0.012	193,381
PATTERN	47,103	0.342	0.013	193,731
Full model			0.108	193,827

Table 3
Regression model performance with β, p value, R^2 and RMSE for predicting real estate selling price. β, p value, R^2 were derived from the training data, RMSE was calculated based on the test data. Values are sorted by RMSE value, in descending order.

Feature	β	p value	R^2	RMSE
NRC	30,014	**0.209**	**0.027**	**175,866**
BING	33,462	0.220	0.026	175,897
VADER	52,349	0.243	0.021	176,458
AFINN	11,533	0.292	0.019	176,615
LIWC15	-6,324	0.385	0.010	176,808
PATTERN	45,819	0.358	0.012	177,111
SentiStrength	8,973	0.374	0.012	177,112
Full model			0.099	178,048

Table 4
Regression model performance with β, p value, R^2 and RMSE for predicting real estate price difference in percentage. β, p value, R^2 were derived from the training data, RMSE was calculated based on the test data. Values are sorted by RMSE value, in descending order.

Feature	β	p value	R^2	RMSE
BING	-0.471	0.489	0.009	**20.401**
AFINN	-0.191	0.515	0.007	20.457
VADER	-1.205	**0.443**	**0.013**	20.457
SentiStrength	0.150	0.484	0.009	20.477
LIWC15	-0.014	0.508	0.009	20.489
NRC	0.091	0.507	0.008	20.561
PATTERN	0.256	0.447	0.011	20.632
Full model			0.064	22.358

The regression evaluation metrics of ask price can be found in table 2, of sold price in table 3 and of price difference in table 4. All the R^2 values for ask price are close to zero, indicating that the model is not able to fit the model properly in order to explain a significant proportion of the variance in real estate asking price. Amongst all models, BING scores the least poorly (R^2 = .029), indicating that the best model fit is able predict 2.9% in the variance in asking price. BING was also able to achieve the lowest prediction errors on the test set (RMSE = 191,956). Pattern.en achieved the worst fit and most errors on the test set (R^2 = .013, RMSE = 193,731). Although NRC was used for baseline purposes, it outperforms all other features when predicting asking price. However, all sentiment features were not able to significantly fit the data (see appendix B, C, D and E for comprehensive summary), since $p > .05$ for all cases, with BING having the lowest p value (p = .203). It is found that of all postal code cases, not all subsets achieved a non-significant result (cases where $p < 0.05$). The percentage of postal cases where results were significant are NRC(48%), BING(46%), VADER(39%), AFINN(35%), Pattern.en (20%), LIWC15(19%) and SentiStrength(18%). An example of one of the cases where the effect was significant, was for the VADER sentiment with postal code 01331. Here, VADER significantly predicted ask price, β = 50,065, $t(142) = 2.58$, $p < .05$. The model achieved a poor fit with R^2 = .03, $F(1, 142) = 6.65$, $p < .05$. The model yielded an RMSE of 75,702.

The R^2 found for the prediction of selling price, are similar to those found for asking price. Sentiment tools achieved the same ranking relative to each other, except for Pattern, which fits less poorly than SentiStrength. Additionally, NRC achieves a better fit and less errors on the test set than BING, making it the best performing model amongst sentiment analysis features for sold price. Significance performance are also similar, with the lowest p value of p = .209 for NRC and highest of p = .374 for SentiStrength. The percentage of significant cases are also similar to those for asking price, with NRC(47%), BING(46%), VADER(38%), AFINN(32%), Pattern.en (21%), SentiStrength(19%) and LIWC15(16%). Another VADER example, but for sold price, also concerns the postal code 01331. VADER significantly predicted sold price, β = 50,693, $t(138) = 2.96$, $p < .05$. The model achieved a poor fit with R^2 = .042, $F(1, 138) = 8.79$, $p < .05$, with an RMSE of 64,562.

For price difference, R^2 values are less favorable than R^2 values for the other two dependent variables. The same holds true for the significance levels for the different features. The highest R^2 value is achieved by VADER (R^2 = .013), as well as for the lowest significance level, p = .443. The lowest RMSE is achieved by BING (RMSE = 20.40). Most noticeably, is that NRC was previously most favorable in predicting asking and selling price but fails to capture the price difference between both. Pattern.en achieved the worst fit and most errors on the test set (R^2 = .013, RMSE = 193,731), below our baseline (NRC). The percentages of postal cases with significance levels of $p < .05$ for price difference were VADER(13%), Pattern.en(11%), LIWC15(8%), SentiStrength(7%), NRC(6%), BING(6%), and AFINN(5%). In contrast to previous dependent variables, the significance levels were present in less cases (13% versus 48% and 47%).

When taking the full models into consideration, in which all sentiment results are used to predict the dependent variables, it was possible to achieve higher model fit for all cases compared to the individual features (R^2_{ask} = .108, R^2_{sold} .099, $R^2_{pricediff}$ = .064). Although a better fit was achieved, higher errors were generated on the test set compared to other features ($RMSE_{ask}$ = 193,827, $RMSE_{sold}$ = 178,048, $RMSE_{pricediff}$ = 22.36).

Table 5
Classification model performance R^2 for predicting real estate ask price, sold price and price difference. F_1 scores indicating Macro-F^1 scores derived from classes 'Decrease' and 'Increase'.

	ASK_CLASS		SOLD_CLASS		DIFF_CLASS		
	SVC	KN	SVC	KN	SVC	KN	Mean F^1
VADER	0.2321	**0.3401**	0.2323	**0.3381**	0.3772	0.6385	**0.3597**
LIWC15	**0.2369**	0.3275	0.2377	0.3249	0.3920	0.6304	0.3582
AFINN	0.2369	0.3284	**0.2391**	0.3259	**0.3943**	0.6110	0.3559
Pattern.en	0.2322	0.3273	0.2326	0.3256	0.3772	**0.6441**	0.3565
NRC	0.2338	0.3358	0.2340	0.3365	0.3820	0.5923	0.3524
BING	0.2326	0.3358	0.2336	0.3331	0.3826	0.5877	0.3509
SentiStrength	0.2323	0.3186	0.2332	0.3216	0.3808	0.5851	0.3453

The evaluation metrics for ask price, sold price and price difference classes are listed in table 3. As with regression, VADER was able to best predict pricing on average (F^1 = .36), by means of classification. Similarly, SentiStrength performed on average the worst amongst the classifiers (F^1 = .35). For each dependent variable, a different combination of sentiment lexicon and classifier yielded the highest prediction metrics. Of the classifiers, KNeighbors outperformed the SVC model on ask price, sold price and price difference (average F^1_{SVC} = .28 and F^1_{KN} = .42). Pattern yielded the highest recorded F^1 value of .64 within all test cases. NRC, our baseline, performed just below average.

5. Discussion

In search for the extend in which sentiment in real estate descriptions is able to predict ask price, sold price and price difference, it was found that prediction using regression modelling yielded low model fit (at most 2.9% was explained with an individual feature). The same holds true for classification modelling when considering ask price and sold price, where the highest F^1 value recorded was .34 by VADER. However, for price difference classification, by using the KNeighbors model, all sentiment lexicons were able to predict 'Decrease and 'Increase' classification (with balanced datasets) above F^1 = .58, with the highest value scored by Pattern.en (F^1 = .64). These findings do set a positive note to this research. For future research, it can be focused to consider classifying with given settings and fine-tune parameters or use more extensive machine/deep learning models. After which, binary classification could be extended by incorporating the 'Equal' class.

An interesting observation one can make, is the fact that two renowned lexicon-based approaches, VADER and SentiStrength, scored on different spectrums when it comes to predictive power for real estate pricing in classification. VADER scored the best results amongst the lexicons, while SentiStrength achieved the lowest. Moreover, VADER scored higher than SentiStrength, even for two class classification. Latter is in contrast with benchmarking performed by Ribeiro et al. (2016), who found that VADER scored highest amongst three class classification (SentiStrength 15th) and SentiStrength the highest for binary classification (VADER 8th).

To address our last formulated research question, whether combining lexicon polarity scores would yield a greater predictive power, it was found that model fit increased, but errors on the testing data increased. We can therefore state that combining the models, does not yield a better predictive performance.

However, this research does have its limitations. First and foremost, is the fact that the model results were aggregated over postal codes, hereby eliminating positive and negative relationships per postal code by taking the overall average. Aggregation was performed to get an overall analysis of the complete data. Significant and non-significant findings were also aggregated and resulted in an overall non-significant dataset. For future research, it can therefore be suggested to do more extensive research towards sentiment in unstructured data of real estate with a more comprehensive and homogenous dataset. Hereby focusing firstly on postal codes in which a significant effect was found, related to the specified sentiment analysis method. By doing so, more specific text and opinion mining can be further examined and combined with hedonic pricing characteristics.

6. Conclusion

Research was performed towards determining the predictive power of sentiment found in real estate descriptions on their related ask price, sold price and price difference. Sentiment was extracted from descriptions found on Zillow.com, a real estate agency based in the US. The sentiment polarity was determined via lexicon-based sentiment analysis methods, more specifically by VADER, AFINN, NRC (EmoLex), LIWC15, BING, Pattern.en and SentiStrength. R^2 Evaluations from the regression models were for all pricing features close to zero, meaning that it was not able to fit the model properly in order to explain a significant proportion of the variance in real estate asking and selling price, nor in price difference. This holds true for all lexicons utilized via a regression model. However, for classification, ask price and sold price the lowest F^1 value measured was .24 by the SVC model, in contrast to .34 by the KNeighbor classifier. For price change classification, the best evaluation metrics were generated by the KNeighbor classifier with F^1 = .64, with similar results across all different sentiment lexicons (minimum value of F^1 = .59 for SentiStrength).

References

Abbasi, A., & Dhar, M. (2014). Benchmarking twitter sentiment analysis tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may. European Language Resources Association (ELRA)*.

Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Lotti, E., Magliani, F., & Manicardi, S. (2016, September). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In *KDWeb*.

Araújo, M., Gonçalves, P., Cha, M., & Benevenuto, F. (2014, April). iFeel: a system that compares and combines sentiment analysis methods. In Proceedings of the 23rd International Conference on World Wide Web (pp. 75-78).

Baviera, T., Sampietro, A., & García-Ull, F. J. (2019). Political conversations on Twitter in a disruptive scenario: The role of "party evangelists" during the 2015 Spanish general elections. *The Communication Review*, *22*(2), 117-138.

Bardhan, A. D, Jaffee, D., & Kroll, C. (2000). The Internet, E-Commerce and the Real Estate Industry. UC Berkeley: Fisher Center for Real Estate and Urban Economics. Retrieved from https://escholarship.org/uc/item/7jx4b9sb

Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, *6*(1), 65-76.

Canuto, S., Gonçalves, M. A., & Benevenuto, F. (2016). Exploiting New Sentiment-Based Meta-level Features for Effective Sentiment Analysis. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16*. doi:10.1145/2835776.2835821

Carvalho, F., Rodrigues, R. G., & Guedes, G. P. (2018, October). LIWBC: a bigram algorithm to enhance results in polarity classification. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web* (pp. 419-422).

Choosing the Right Estimator. scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

Culpeper, J., Findlay, A., Cortese, B., & Thelwall, M. (2018). Measuring emotional temperatures in Shakespeare's drama. English Text Construction, 11(1), 10-37.

Das, S., & Chen, M. (2001, July). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific finance association annual conference (APFA)* (Vol. 35, p. 43).

De Smedt, T., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, *13*(66), 2063-2067.

del Pilar Salas-Zárate, M., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., & Alor-Hernández, G. (2014). A study on LIWC categories for opinion mining in Spanish reviews. *Journal of Information Science*, *40*(6), 749-760.

Dezember, R., & Rudegeair, P. (2019, June 19). The Future of Housing Rises in Phoenix. Retrieved June 11, 2020, from https://www.wsj.com/articles/the-future-of-housing-rises-in-phoenix-11560957036

Ecommerce Market Share from 2017 to 2023. (2020, April 23). Retrieved June 10, 2020, from https://www.oberlo.com/statistics/ecommerce-share-of-retail-sales

eCommerce - worldwide: Statista Market Forecast. (2020). Retrieved June 10, 2020, from https://www.statista.com/outlook/243/100/ecommerce/worldwide

Enevoldsen, K. C., & Hansen, L. (2017). Analysing political biases in danish newspapers using sentiment analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, *2*(2), 87-98.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. IEEE Transactions on Knowledge and Data Engineering, 23(10), 1498-1512. doi:10.1109/tkde.2010.188

Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013, October). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks* (pp. 27-38).

Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In *AAAI* (Vol. 4, No. 4, pp. 755-760).

Huang, P., Lurie, N. H., & Mitra, S. (2009). Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods. Journal of Marketing,73(2), 55-69. doi:10.1509/jmkg.73.2.55

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Hutto, C. J. (2020). Cjhutto/vaderSentiment. Retrieved June 19, 2020, from https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vaderSentiment.py

Islam, M. R., & Zibran, M. F. (2018). SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of Systems and Software, 145*, 125-146. doi:10.1016/j.jss.2018.08.030

Jockers, M. (2017). Package 'syuzhet'. URL: https://cran. r-project. org/web/packages/syuzhet.

Jockers, M. (2019, May 11). Mjockers/syuzhet. Retrieved June 20, 2020, from https://github.com/mjockers/syuzhet

Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, *11*(3), 205-217.

Lee, E., & Shin, S. Y. (2014). When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo. Computers in Human Behavior, 31, 356-366. doi:10.1016/j.chb.2013.10.050

Loria, S. (2017). Sloria/TextBlob. Retrieved June 19, 2020, from https://github.com/sloria/TextBlob/blob/dev/textblob/en/sentiments.py

Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on Worldwide Web* (pp. 342-351).

Ling, D. C., Ooi, J. T., & Le, T. T. (2015). Explaining house price dynamics: Isolating the role of nonfundamentals. Journal of Money, Credit and Banking, 47(S1), 87-125.

Gan, Q., & Yu, Y. (2015, January). Restaurant Rating: Industrial Standard and Word-of-Mouth--A Text Mining and Multi-dimensional Sentiment Analysis. In *2015 48th Hawaii International Conference on System Sciences* (pp. 1332-1340). IEEE.

Jindal, N., & Liu, B. (2006, August). Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 244-251).

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of
six sentiment lexicons. *Journal of Information Science*, *44*(4), 491-511.

Mohammad, S., & Turney, P. (2010, June). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 26-34).

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*(3), 436-465.

Mohammad, S. (2018). Accessing the NRC Emotion and Sentiment Lexicons. Retrieved June 20, 2020, from http://saifmohammad.com/WebPages/AccessResource.htm

Mohammad, S. (2020). NRC Word-Emotion Association Lexicon (aka EmoLex). Retrieved June 20, 2020, from http://saifmohammad.com/WebPages/AccessResource.htm

Mou, Jian, et al. "Impact of Product Description and Involvement on Purchase Intention in Cross-Border e-Commerce." *Industrial Management & Data Systems*, vol. 120, no. 3, 2019, pp. 567–586., doi:10.1108/imds-05-2019-0280.

Naldi, M. (2019). A review of sentiment computation methods with R packages. *arXiv preprint arXiv:1901.08319*.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Nielsen, F. (2017, August 21). Fnielsen/afinn. Retrieved June 20, 2020, from https://github.com/fnielsen/afinn/blob/master/afinn/afinn.py

NLTK Project. (2020, April 13). Natural Language Toolkit. Retrieved June 18, 2020, from https://www.nltk.org/

NLTK. (2020). Source code for nltk.sentiment.vader. Retrieved June 16, 2020, from https://www.nltk.org/_modules/nltk/sentiment/vader.html

Orasan, C. (2018, June). Aggressive language identification using word embeddings and sentiment features. Association for Computational Linguistics.

Owusu-Ansah, A. (2011). A review of hedonic pricing models in housing research. *Journal of International Real Estate and Construction Studies*, *1*(1), 19.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science, 8,* 162166.

Pennebaker, J. W. (2002). What our words can say about us: Towards a broader language psychology. *Psychological Science Agenda, 15,* 89.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count (LIWC): LIWC2007.* Austin, TX: LIWC.net.

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin.

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. *Journal of real estate research*, *31*(2), 147-164.

Piolat, A., Booth, R., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). The French dictionary for LIWC: Modalities of construction and examples of use| La version franaise du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, 5(1), 1-29.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., & Stoyanov, V. (2015). SemEval-2015 Task 10: Sentiment Analysis in Twitter. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. doi:10.18653/v1/s15-2078

Sadia, A. (2016). Sentiment Analysis Using Language Model. *Asian Journal of Engineering,] Sciences & Technology*.

Sadia, A., Khan, F., & Bashir, F. (2018). An Overview of Lexicon-Based Approach for Sentiment Analysis.

Sarkar, D. (2019). Text Analytics with Python. doi:10.1007/978-1-4842-4354-1

Schmedt, T. D. (2014, May 10). Clips/pattern. Retrieved June 19, 2020, from

https://github.com/clips/pattern/blob/master/pattern/text/en/en-sentiment.xml

SentiStrength. (n.d.). Retrieved June 19, 2020, from http://sentistrength.wlv.ac.uk/

Steven, Dick. "Predicting Real Estate Price Using Text Mining." Tilburg University, 2014, arno.uvt.nl/show.cgi?fid=134740.

Seih, Y. T., Beier, S., & Pennebaker, J. W. (2017). Development and examination of the linguistic category model in a computerized text analysis method. Journal of Language and Social Psychology, 36(3), 343-355.

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of real estate literature*, *13*(1), 1-44.

Singh, V., Singh, G., Rastogi, P., & Deswal, D. (2018). Sentiment Analysis Using Lexicon Based Approach. *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. doi:10.1109/pdgc.2018.8745971

Thelwall, M. (2018). Gender bias in sentiment analysis. *Online Information Review*.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544-2558. doi:10.1002/asi.21416

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163-173.

Tong, R. M. (2001, September). An operational system for detecting and tracking opinions in on-line discussion. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* (Vol. 1, No. 6).

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., & Daelemans, W. (2016). A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010, May). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*.

Urolagin, S. (2017, September). Text mining of tweet for sentiment classification and association with stock prices. In *2017 International Conference on Computer and Applications (ICCA)* (pp. 384-388). IEEE.

Wang, S., Lv, G., Mazumder, S., Fei, G., & Liu, B. (2018, December). Lifelong learning memory networks for aspect sentiment classification. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 861-870). IEEE.

Vural, A. G., Cambazoglu, B. B., Senkul, P., & Tokgoz, Z. O. (2013). A framework for sentiment analysis in Turkish: Application to polarity detection of movie reviews in Turkish. In *Computer and Information Sciences III* (pp. 437-445). Springer, London.

Vilares, D., Thelwall, M., & Alonso, M. A. (2015). The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *Journal of Information Science*, *41*(6), 799-813.

Zhao, N., Jiao, D., Bai, S., & Zhu, T. (2016). Evaluating the validity of simplified Chinese version of LIWC in detecting psychological expressions in short texts on social network services. *PloS one*, *11*(6).

Zillow Group. (2020). *2019 Annual Report and Form 10K*. Retrieved from http://www.annualreports.com/HostedData/AnnualReports/PDF/NASDAQ_Z_2019.pdf

Zillow, Inc. (2019, June). What is a Zestimate? Zillow's Zestimate Accuracy. Retrieved June 12, 2020, from https://www.zillow.com/zestimate/

Zillow, Inc. (2019, November 27). Property Descriptions 101: How to Write Listing Descriptions That Sell: Zillow. Retrieved from https://www.zillow.com/sellers-guide/listing-descriptions-that-sell/

7. Appendix A: Raw Zillow data

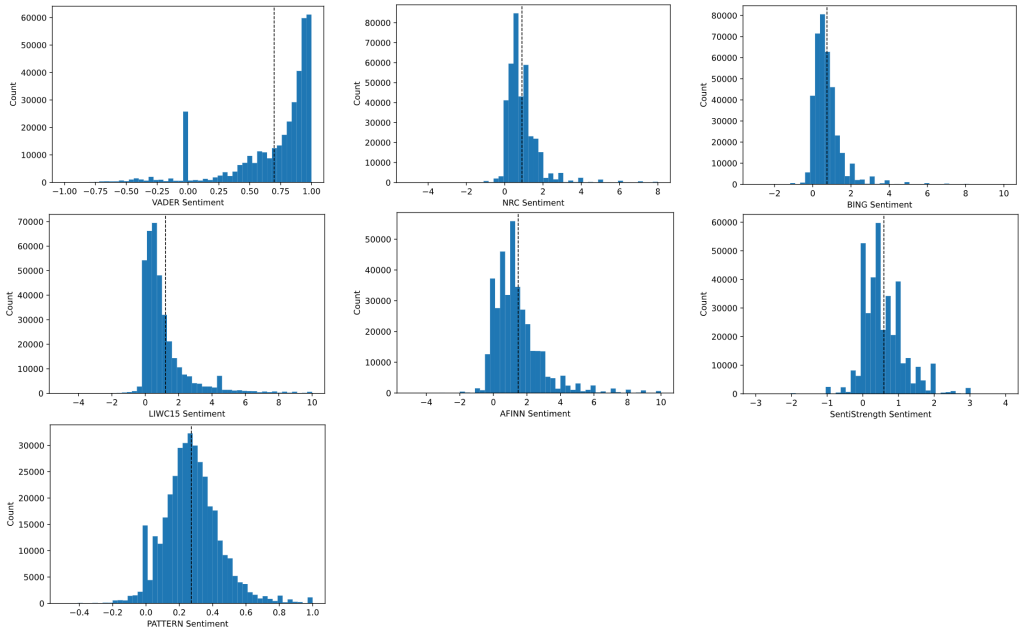
Table 5: Raw data

ID	'00852-1',
Price	'$190,000',
Sold price	'$190,000',
City State	Maple Shade, NJ 08052
$_sqft	'$103',
Sqft	'1,840',
Type	'Single Family',
Year_built	'1912',
Street	'103 W Linwood Ave',
Extended_adress	'--',
Sale duration	95.0,
Baths	'2',
Beds	'3',
Cooling	'--',
Lot'	'7,405 sq ft / 0.17 acres',
Fireplace	'--',
Furnished:	'--',
Heating	'--',
Disability access	'--',
Utilities included	'--',
High-speed internet ready	'--',
Onsite laundry	'--',
Parking	'--',
Pets	'--',
Availability	'--',
Lease term	'--',
Listing website	'--',
Photodir	'--',
Photos	'--',
Property website	'--',
MLS_#	'--',
On_Zillow	'--',

History	'Date @@@ 09/10/2012\tDescription @@@ Sold\tPrice @@@ $190,000\tChange @@@ -5.0%\t$/sqft @@@ $103\tSource @@@ Public Record\tDate @@@ 07/31/2012\tDescription @@@ Listing removed\tPrice @@@ $199,999\tChange @@@ --\t$/sqft @@@ $108\tSource @@@ CENTURY 21 Alliance\tDate @@@ 06/20/2012\tDescription @@@ Pending sale\tPrice @@@ $199,999\tChange @@@ --\t$/sqft @@@ $108\tSource @@@ CENTURY 21 Alliance\tDate @@@ 06/07/2012\tDescription @@@ Listed for sale\tPrice @@@ $199,999\tChange @@@ 590%\t$/sqft @@@ $108\tSource @@@ CENTURY 21 Alliance\tDate @@@ 08/14/1992\tDescription @@@ Sold\tPrice @@@ $29,000\tChange @@@ --\t$/sqft @@@ $15\tSource @@@ Public Record\t'
Description	"This is a 3 bedroom 2 bath Bungalow Remodel you don't want to pass by! I promise, you will not be disappointed unless you don't see it! Gorgeous Redone Kitchen with beautiful Granite countertops. All neutral colors, Hardwood floors throughout except Bedrooms are nicely carpeted. Entire upstairs for Master suite and full bath. Living room and Dining room Remodeled, nicely formal with Crown molding and trim around the doors and windows. The living room has a nice white built-in bookshelf. I just cannot say enough about the beautiful workmanship that has gone into this home and the care with which it has been maintained. The roof is also above the rest with a 50 year warranty that goes with the sale of the house! The corner lot has a fantastic garden for herbs and veggies! Bright Sunny Happy home to start your own family. Close to everything even if you work in Phiily! Blocks from the Quaint downtown Maple Shade Shopping, and community festivities! Great home to entertain on Holidays! TY for comin you'll luv me!",

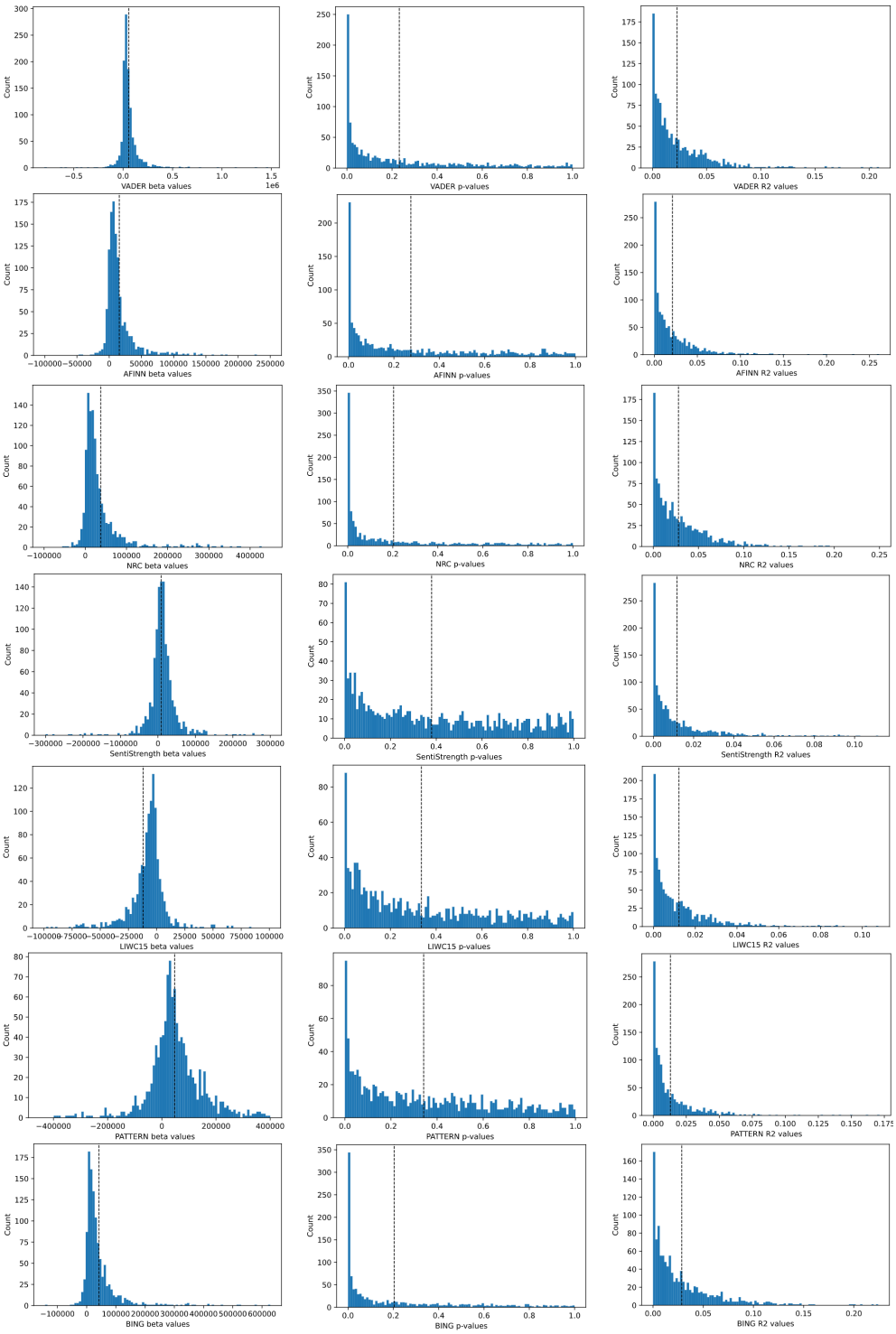
8. Appendix B: Results Sentiment Analysis Methods

Figure 6
Overview of the seven sentiment analysis distributions.



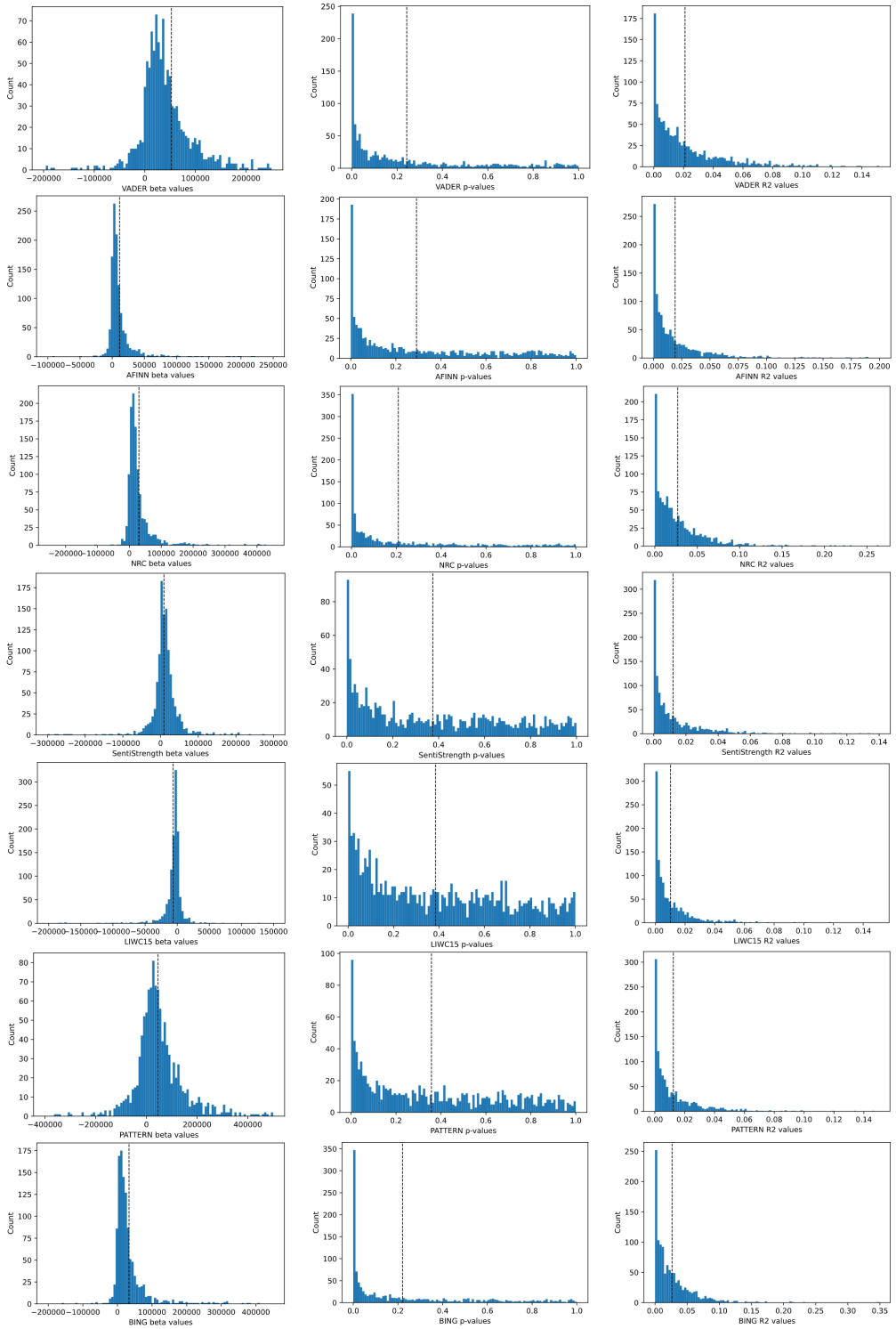
9. Appendix C: Results Evaluation Metrics for Ask Price

Figure 7
Overview of the evaluation metrics for the seven sentiments on ask price.



10.Appendix D: Results Evaluation Metrics for Sold Price

Figure 8
Overview of the evaluation metrics for the seven sentiments on sold price.



11.Appendix E: Results Evaluation Metrics for Price Difference

Figure 9
Overview of the evaluation metrics for the seven sentiments on price difference.