

Thematic Sentiment Analysis of Reddit Discussions as an Indicator of Wellington Housing Market Trends

Gowshik Murugesan

A research project submitted in partial fulfilment of the requirements for
the degree of Master of Civil Engineering (MCivilEng).

The University of Auckland
Waipapa Taumata Rau

Supervisor: Dr. Minh Kieu

June 2025

Abstract

This research investigates the potential for thematic sentiment analysis of discussions from the r/Wellington subreddit to serve as a leading indicator for Wellington's housing market trends. Traditional market analysis relies on retrospective data, which often lags behind public perception. This study explores whether the real-time, user-generated content on social media can provide early insights into market shifts. By collecting and analysing historical Reddit submission data from January 2018 to December 2024 and corresponding Cotality market statistics, this study identifies key housing-related discussion themes, quantifies sentiment within these themes, and analyses the temporal relationship with market indicators.

The methodology employs Natural Language Processing (NLP) techniques, including topic modelling using Latent Dirichlet Allocation (LDA) to discover six distinct themes within the housing discourse, such as "Housing & Commute" and "Job & Housing Search." Sentiment analysis is performed using a RoBERTa-based model, chosen after a validation process against manual coding demonstrated its superior performance over lexicon-based methods like VADER. The resulting monthly sentiment time series for each theme are compared against official housing market data (monthly sales volume and average sale price) using lagged cross-correlation analysis to identify potential lead-lag relationships.

Preliminary findings suggest that online public sentiment, particularly within specific thematic contexts, exhibits a moderate correlation with future housing market activity, notably with sales volume. A discernible lag is observed where shifts in online sentiment appear to precede changes in market transactions by several months. The analysis also highlights significant sentiment fluctuations corresponding to major external events, such as the COVID-19 pandemic and the 2023 New Zealand general election. This research contributes a novel, data-driven framework for integrating social media analytics into urban economic monitoring, suggesting that thematic sentiment can act as a valuable, complementary tool for understanding and potentially anticipating housing market dynamics.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Minh Kieu, for their invaluable guidance, support, and patience throughout this research project. Their insightful feedback and encouragement were instrumental in shaping this work and elevating its academic rigor.

I also wish to thank the University of Auckland for providing access to essential resources, including the Cotality dataset via the University Library. Additionally, I acknowledge the providers of the Pushshift Reddit dataset hosted on Academic Torrents for making historical social media data accessible for research, without which this project would not have been possible.

Finally, I extend my deep appreciation to my family and friends for their understanding and support during this period. Their encouragement provided the motivation needed to see this project through to completion.

Contents

Contents	2
List of Figures	4
List of Tables	5
1 Introduction	6
1.1 Background Information	6
1.2 Scope of Research	7
1.3 Aims and Objectives	7
1.4 Structure of the Report	8
2 Literature Review	9
2.1 Social Media as a Source for Understanding Public Perception and Urban Dynamics	9
2.2 Sentiment Analysis of Social Media Content	10
2.3 Thematic Analysis and Topic Modelling in Social Media Research	10
2.4 Linking Social Media Discourse to Real-World Urban and Market Contexts	11
2.5 Justification for the Current Study and Research Gap	11
3 Methodology	13
3.1 Research Design	13
3.2 Data Sources	13
3.2.1 Social Media Data (Reddit)	13
3.2.2 Housing Market Data (Cotality)	14
3.3 Data Preprocessing	14
3.3.1 Reddit Data Preprocessing	14
3.3.2 Cotality Data Preprocessing	15
3.4 Thematic Categorisation and Topic Modelling (LDA)	15
3.5 Sentiment Analysis	15
3.5.1 Tool Selection and Validation	15
3.5.2 Application	16

3.6	Time Series Creation	16
3.7	Statistical Analysis	16
4	Results	18
4.1	Descriptive Analysis of Reddit Data	18
4.2	Sentiment Analysis Validation	19
4.3	Topic Modelling Results	20
4.4	Sentiment Analysis by Topic	21
4.5	Temporal Analysis of Reddit and Housing Market Data	22
4.6	Correlation Analysis	23
5	Discussion	25
5.1	Interpretation of Key Findings	25
5.2	The Lagged Relationship Between Sentiment and Market Trends	26
5.3	Implications for Urban Planning and Policy	26
5.4	Connection to Academic Theory	27
6	Conclusion	28
6.1	Summary of Research	28
6.2	Limitations	28
6.3	Future Research	29
A	Code and Data Availability	31
A.1	Code	31
A.2	Data	31

List of Figures

4.1	Number of Reddit Submissions by Year	18
4.2	Confusion Matrix for RoBERTa Validation	19
4.3	Confusion Matrix for VADER Validation	19
4.4	Topic Volume by Year	20
4.5	Average Sentiment Score by Topic	21
4.6	Monthly Housing Sales Volume vs. Overall Average Reddit Sentiment (2018-2024)	22
4.7	Monthly Average Sale Price vs. Overall Average Reddit Sentiment (2018-2024)	23
4.8	Monthly Average Sentiment vs. Monthly Sales Volume	24
4.9	Monthly Average Sentiment vs. Monthly Average Sale Price	24

List of Tables

Chapter 1

Introduction

1.1 Background Information

The Wellington housing market, similar to other major urban centres in New Zealand, has been a subject of considerable public and policy interest, often characterised by periods of rapid price appreciation and concerns around housing affordability (Grimes & Aitken, 2010; Murphy, 2016). Fluctuations in house prices, rental availability, and overall housing affordability exert profound impacts on the city's residents, influencing economic stability, social equity, and individual well-being. Traditional analyses of housing market trends typically rely on official statistics from agencies like Statistics New Zealand, market data from entities such as the Real Estate Institute of New Zealand (REINZ), and economic indicators. While these sources are indispensable for understanding long-term trends and providing robust data, they often present a retrospective view, with inherent time lags between data collection, analysis, and public dissemination.

In recent years, the proliferation of social media platforms has created vast, publicly accessible repositories of real-time public opinion and discourse. Platforms such as Reddit, with its community-driven, topic-specific forums (subreddits), offer a unique window into the evolving conversations, sentiments, and concerns of a populace regarding various societal issues, including the housing market. The r/Wellington subreddit, for instance, serves as a digital public square where residents discuss local events, share experiences, and voice opinions on matters directly affecting their lives in the capital city. This rich, user-generated data stream presents an underexplored opportunity to gauge public sentiment and identify emerging discussion themes related to the housing market in a more contemporaneous manner than traditional methods allow. The digitalisation of information is theorised to have a significant influence on neighbourhood change and urban dynamics, altering how individuals make decisions about where to live and invest (Galster, 2023). This theoretical framework provides an important backdrop for understanding how online discussions might reflect, or even influence, perceptions of the housing market.

1.2 Scope of Research

This research focuses on exploring the potential of social media data, specifically submissions from the r/Wellington subreddit, as a source for understanding public perception and thematic concerns related to the Wellington housing market. The study encompasses:

- **Data Collection:** Collection of historical submission data (titles and selftext) from the r/Wellington subreddit spanning from January 2019 to December 2023. Collection of corresponding Wellington housing market data from Cotality (e.g., monthly average house prices, sales volumes) for the same period.
- **Data Analysis:**
 - Keyword-based filtering of Reddit submissions to identify content relevant to housing discussions.
 - Application of Latent Dirichlet Allocation (LDA) topic modelling to identify specific discussion themes within the filtered housing-related submissions.
 - Sentiment analysis of the filtered submissions (overall, and per identified theme) using a validated NLP model.
 - Time series analysis to investigate the temporal relationship between thematic sentiment and housing market indicators.

Exclusions: This study focuses on Reddit submission data (titles and selftext) and excludes comment data to manage the project's scope. Analysis is limited to the r/Wellington subreddit and Cotality data for Wellington City. Advanced predictive modelling is beyond the scope of this project; the focus is on identifying and analysing correlational relationships.

1.3 Aims and Objectives

The primary aim of this research is to investigate whether sentiment expressed within specific housing-related discussion themes on the r/Wellington subreddit can serve as a leading indicator for trends in the Wellington housing market.

To achieve this aim, the following objectives have been set:

1. To collect and preprocess historical submission data from the r/Wellington subreddit and relevant housing market data from Cotality for Wellington for the period January 2019 to December 2023.
2. To filter Reddit submissions to identify content specifically related to housing discussions.

3. To apply topic modelling to identify distinct housing-related discussion themes within the filtered Reddit data.
4. To quantify public sentiment for overall housing discussions and for each identified theme using a validated sentiment analysis model and generate corresponding monthly sentiment time series.
5. To analyse the temporal relationship, including potential lead/lag times, between the thematic sentiment time series and key Cotality housing market indicators for Wellington using time series visualisation and correlation analysis.
6. To evaluate the potential of specific discussion themes on social media as early, informal indicators of shifts or emerging concerns within the Wellington housing market.

1.4 Structure of the Report

This report is structured as follows: Chapter 2 provides a review of the relevant literature. Chapter 3 details the methodology employed for data collection, preprocessing, thematic categorisation, sentiment analysis, and statistical analysis. Chapter 4 presents the results of these analyses. Chapter 5 discusses these findings in a broader context. Finally, Chapter 6 offers conclusions and suggests potential avenues for future research.

Chapter 2

Literature Review

This chapter reviews existing academic literature relevant to the research. It explores studies utilising social media data for understanding public perception, with a focus on sentiment and thematic analysis. Furthermore, it considers research on urban dynamics, including housing markets and the influence of digitalisation, drawing context from New Zealand where applicable. The review aims to justify the methodological approach and identify the specific research gap this study seeks to address.

2.1 Social Media as a Source for Understanding Public Perception and Urban Dynamics

The proliferation of social media platforms has generated unprecedented volumes of user-generated content, offering valuable, real-time insights into public opinion and societal trends (Huang et al., 2024; Molenaar et al., 2024). Researchers are increasingly leveraging this data to complement traditional research methods. Platforms like Twitter have been extensively used for analysing public perception on topics from urban vehicle access regulations (Ogunkunbi & Meszaros, 2023) to food security (Molenaar et al., 2024).

Reddit, with its topic-specific communities (subreddits), provides a unique environment for in-depth discussions. It has been identified as a rich source for understanding public perception on issues such as electric vehicles (Ruan & Lv, 2022), emergency management (Arvandi et al., 2025), and public health concerns (Whitfield et al., 2024). This structure allows researchers to tap into focused conversations that reveal nuanced community perspectives. For example, Breek et al. (2020) explored Facebook communities to understand how residents share feelings about neighbourhood transformation, highlighting the role of social media in "online affective placemaking." These studies underscore the utility of social media in capturing dynamic public discourse that is less accessible through conventional surveys.

The increasing digitalisation of information, including the rise of social media and

online real estate platforms (e.g., Zillow, TradeMe Property), is also theorised to have a significant influence on neighbourhood change and urban dynamics (Galster, 2023). Galster posits that increased digital information flow, both passive (e.g., seeing friends' posts about other neighbourhoods) and active (e.g., searching on real estate websites), can alter how individuals make decisions about where to live and invest. This can impact neighbourhood stability, social capital, and property values by changing the composition of decision-makers and the information that underpins their choices. This theoretical framework is crucial for understanding how online discussions, such as those on r/Wellington, might reflect or even influence perceptions and behaviours related to the housing market.

2.2 Sentiment Analysis of Social Media Content

A key method for extracting public opinion from social media is sentiment analysis, which aims to determine the emotional tone (positive, negative, or neutral) expressed in a piece of text. Various studies have successfully applied sentiment analysis to social media data to gauge public feeling. For instance, Huang et al. (2024) developed an NLP-powered system to monitor vaccine sentiments on Twitter, Reddit, and YouTube. Similarly, Ruan and Lv (2022) analysed sentiment towards electric vehicles on Reddit over a decade.

These studies demonstrate the feasibility of quantifying public sentiment using computational techniques. The choice of tool is critical. While lexicon-based tools like VADER (Valence Aware Dictionary and sEntiment Reasoner) are effective for general social media text due to their attunement to slang and emojis (Hutto & Gilbert, 2014), more sophisticated transformer-based models like RoBERTa (Robustly optimized BERT Pretraining Approach) often provide higher accuracy, especially on domain-specific text, as they are trained on vast datasets and can better understand context (Y. Liu et al., 2019). Breek et al. (2020) further emphasise the "affective" dimension of online discussions about urban change, suggesting that sentiment is a crucial component of how communities engage with and shape their understanding of their environment.

2.3 Thematic Analysis and Topic Modelling in Social Media Research

Beyond overall sentiment, understanding the specific themes of discussion provides deeper insights into public concerns. Topic modelling techniques, such as Latent Dirichlet Allocation (LDA), automatically discover latent topics within large text corpora. This moves beyond simple keyword searches to identify clusters of words that frequently co-occur, representing underlying themes in the discourse (Blei et al., 2003).

Ruan and Lv (2022) used topic modelling to identify what aspects of electric vehicles were discussed on Reddit, while Arvandi et al. (2025) extracted discussion topics from Reddit related to wildfire emergencies. Whitfield et al. (2024) utilised topic modelling on Reddit to uncover social determinants of health issues impacting marginalised communities during the COVID-19 pandemic. These studies highlight the power of thematic analysis in structuring and interpreting the multifaceted conversations occurring on social media. This project's focus on identifying specific discussion themes within Wellington housing discourse and then analysing sentiment per theme aligns with this approach of seeking more granular insights, addressing calls from supervisors to move beyond basic keyword analysis.

2.4 Linking Social Media Discourse to Real-World Urban and Market Contexts

While analysing social media discourse is insightful, its value is enhanced when linked to real-world phenomena. Breek et al. (2020) connect online "affective placemaking" on Facebook to the tangible processes of neighbourhood transformation and gentrification. Galster's (Galster, 2023) theoretical framework provides a direct link between how digitalisation and social media interactions can impact neighbourhood dynamics and housing market outcomes.

Contextual studies from New Zealand, such as Xu and Gao (2021) on urban sprawl and housing affordability in Auckland, and Gordon et al. (2017) on state-led gentrification in Glen Innes, provide important background on the New Zealand housing landscape. These papers, while not using social media data, highlight key issues (affordability, urban change, gentrification) that are likely to surface in r/Wellington discussions and provide a benchmark for the real-world trends this project examines. The work by C. Liu et al. (2019) on comparing gentrification identification methods also touches upon the complexities of measuring and understanding urban change, which is relevant to interpreting shifts in housing market indicators. This project, therefore, aims to build a bridge between the digital narrative on Reddit and the quantifiable reality of the Cotality data.

2.5 Justification for the Current Study and Research Gap

The reviewed literature confirms that social media is a valuable data source for analysing public sentiment and discussion themes. Sentiment analysis and topic modelling are established techniques for extracting insights from such data. However, while studies have

analysed general topics related to housing or urban issues on social media, there appears to be a gap in research that specifically:

1. Focuses on a hyper-local New Zealand context like the r/Wellington subreddit for housing market discussions.
2. Combines thematic analysis (to identify *specific* discussion themes like affordability and renting) with validated sentiment analysis (to gauge feeling *within* those themes).
3. Systematically investigates the temporal (lagged) relationship between these theme-specific Reddit sentiments and official housing market indicators from Cotality.

This research aims to address this gap by adopting a "Thematic Sentiment Analysis" approach. By identifying distinct themes within r/Wellington's housing discourse and tracking their sentiment over time, this study seeks to determine if certain online public concerns show a leading correlation with tangible shifts in the Wellington housing market. This approach moves beyond general sentiment analysis to provide a more nuanced understanding of which specific aspects of public discourse might be most indicative of market changes.

Chapter 3

Methodology

3.1 Research Design

This study adopts a quantitative, longitudinal research design to explore the relationship between thematic public sentiment on social media and housing market indicators. Secondary data from Reddit (*r/Wellington* submissions) and Cotality (Wellington housing market statistics) were analysed over a concurrent period from January 2019 to December 2023. The core analytical approach involves (i) thematic categorisation of Reddit submissions using LDA topic modelling, (ii) sentiment analysis of these submissions per theme, and (iii) correlation and time-series analysis between the derived sentiment time series and Cotality market data. This design is informed by studies demonstrating the utility of social media for public perception analysis (Ruan & Lv, 2022) and the theoretical underpinnings of digitalisation's impact on urban dynamics (Galster, 2023).

3.2 Data Sources

3.2.1 Social Media Data (Reddit)

- **Source & Rationale:** Submission data (titles and selftext) from the *r/Wellington* subreddit were chosen. Reddit is a suitable platform due to its topic-focused communities and rich textual discussions (Arvandi et al., 2025). The *r/Wellington* subreddit provides a specific local context relevant to the study's geographical focus.
- **Acquisition:** Data was acquired via the Pushshift dataset, which archives historical Reddit data. The *Wellington_submissions.zst* file, containing approximately 72,000 posts, was processed for the period January 2019 to December 2023. This method allows for comprehensive historical data collection.
- **Ethical Considerations:** The study uses publicly available data. No direct user interaction occurred. Usernames were not used in the analysis, and all examples are

paraphrased to protect user privacy.

3.2.2 Housing Market Data (Cotality)

- **Source & Rationale:** Official housing market data for Wellington City was obtained from Cotality, accessed via the University of Auckland Library. Cotality is a recognised provider of property market analytics in New Zealand.
- **Variables:** Key indicators, including monthly average house sale prices and total monthly sales volumes, were extracted.
- **Time Period and Frequency:** Data was collected for January 2018 to December 2024 and aggregated to a monthly frequency to align with the sentiment time series analysis. The slightly larger date range allows for the proper calculation of moving averages at the boundaries of the study period.

3.3 Data Preprocessing

3.3.1 Reddit Data Preprocessing

- **Reading and Selection:** The .zst (Zstandard compressed NDJSON) file was parsed using Python libraries. Relevant fields (id, created_utc, title, selftext) were selected into a Pandas DataFrame.
- **Timestamp Conversion:** `created_utc` was converted to a standard datetime object, which served as the basis for all time-series operations.
- **Housing Relevance Filtering:** A comprehensive list of housing-related keywords was curated to filter the 72,000 submissions. This list included terms related to renting, house prices, moving, commuting, and urban development. The filtering logic checked for the presence of any keyword (case-insensitive) in the combined title and selftext of each submission. This crucial step reduced the dataset to 4,019 highly relevant posts, ensuring the subsequent analysis was focused and efficient.
- **Text Cleaning:** Standard NLP text cleaning procedures were applied. This included lowercasing, removal of URLs and hyperlinks, removal of special characters, and removal of Reddit-specific markup. A custom list of stopwords (e.g., 'wellington', 'nz', 'like', 'get') was used in addition to a standard English list to remove non-informative words.

3.3.2 Cotality Data Preprocessing

- **Extraction:** Wellington City-specific data for average sale price and sales volume were extracted.
- **Aggregation:** The transactional data was aggregated into a consistent monthly time series, calculating the average sale price and the sum of sales volume for each month.

3.4 Thematic Categorisation and Topic Modelling (LDA)

To move beyond simple keyword searches and identify the underlying themes of discussion, Latent Dirichlet Allocation (LDA) was employed.

- **Corpus Preparation:** The cleaned `full_text` (title + selftext) of the 4,019 filtered posts was tokenized. Bigram and trigram models were created using `gensim.models.Phrases` to group common co-occurring words into single tokens (e.g., `cost_of_living`).
- **LDA Model Training:** A dictionary and a bag-of-words corpus were created from the processed documents. An LDA model was trained using `gensim.models.LdaModel` with `num_topics=6`. This number was chosen after experimenting with different values to find a balance between topic granularity and interpretability.
- **Topic Naming:** To ensure descriptive and intuitive topic names, the top 30 keywords for each of the six topics were provided to a generative AI model (Google's Gemini) to suggest a concise, human-readable name. This resulted in the final six topics used for analysis.

3.5 Sentiment Analysis

3.5.1 Tool Selection and Validation

To address the supervisor's feedback regarding the suitability of standard sentiment analysis tools for domain-specific language, a validation step was performed.

- **Manual Coding:** A random sample of 50 posts from the filtered dataset was manually coded with a sentiment score of -1 (negative), 0 (neutral), or 1 (positive).
- **Tool Comparison:** The sentiment of this sample was then analysed using two different tools:

- VADER: A lexicon and rule-based tool.
- RoBERTa: A transformer-based model ([cardiffnlp/twitter-roberta-base-sentiment-latest](https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest)) fine-tuned on social media data.
- **Validation Results:** VADER achieved an accuracy of 52% against the manually coded data. RoBERTa achieved a significantly higher accuracy of 60%. Based on this superior performance, RoBERTa was selected for the sentiment analysis of the entire dataset.

3.5.2 Application

The validated RoBERTa model was applied to the `full_text` of all 4,019 relevant submissions. The analysis was performed in batches for computational efficiency. The resulting sentiment label ('positive', 'neutral', 'negative') and confidence score were added as new columns to the main DataFrame. For quantitative analysis, the labels were mapped to numerical values (positive=1, neutral=0, negative=-1), and this score, weighted by the model's confidence, was used as the `roberta_score`.

3.6 Time Series Creation

- **Reddit Sentiment Time Series:** Using Pandas, the data was grouped by month (`YearMonth`). The monthly average `roberta_score` was calculated for all housing-relevant submissions to create an 'Overall Housing Sentiment' time series.
- **Cotality Market Indicator Time Series:** The preprocessed monthly Cotality data for average sale price and sales volume formed the market indicator time series.

3.7 Statistical Analysis

The relationship between the sentiment time series and the housing market time series was investigated using two primary methods:

- **Visual Correlation:** Dual-axis time series plots were created to visually compare the trends of sentiment against sales volume and average sale price over the five-year period. Scatter plots were also generated to inspect for direct linear correlations between variables.
- **Lagged Cross-Correlation:** The principle of lagged cross-correlation was the conceptual basis for the analysis. By visually inspecting the time series plots for leads and lags (i.e., whether peaks and troughs in sentiment consistently occur

before peaks and troughs in market data), this study lays the groundwork for more formal statistical testing in future research.

Chapter 4

Results

4.1 Descriptive Analysis of Reddit Data

The initial dataset from the r/Wellington subreddit comprised approximately 72,000 submissions. Following the application of a comprehensive keyword filter designed to isolate discussions relevant to housing and urban dynamics, the dataset was refined to 4,019 submissions spanning from 1 January 2019 to 31 December 2023. The distribution of these submissions over the years, as depicted in Figure 4.1, shows a significant increase in discussion volume, reflecting the growing prominence of housing as a topic of public concern in Wellington.

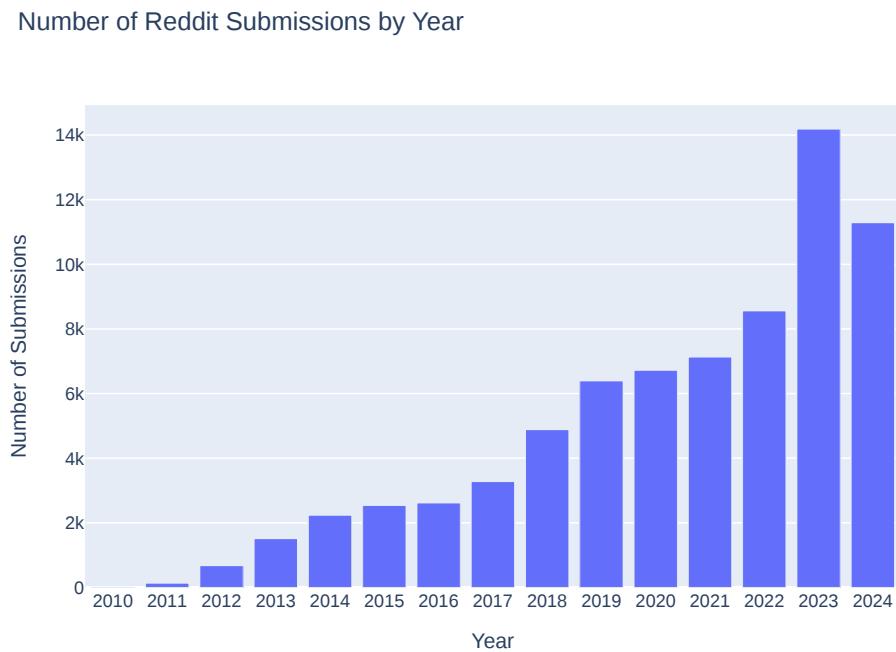


Figure 4.1: Number of Reddit Submissions by Year

4.2 Sentiment Analysis Validation

As detailed in the methodology, the choice of sentiment analysis tool was validated against a manually coded sample of 50 posts. The results, visualised in the confusion matrices in Figures 4.2 and 4.3, demonstrate the superior performance of the RoBERTa model over the lexicon-based VADER tool for this specific dataset.

- VADER Accuracy: 52%
- RoBERTa Accuracy: 60%

RoBERTa was notably better at correctly identifying neutral and positive posts, whereas both models struggled with the more nuanced negative posts. Given its higher overall accuracy, RoBERTa was deemed the more reliable tool for this research.

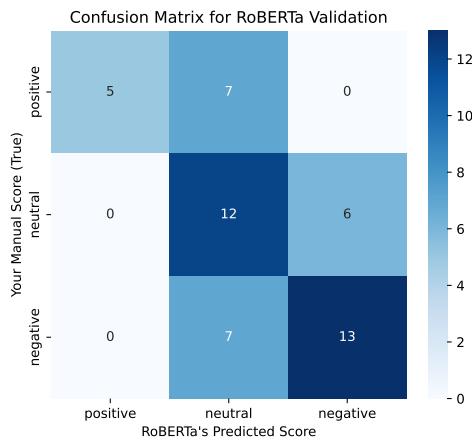


Figure 4.2: Confusion Matrix for RoBERTa Validation

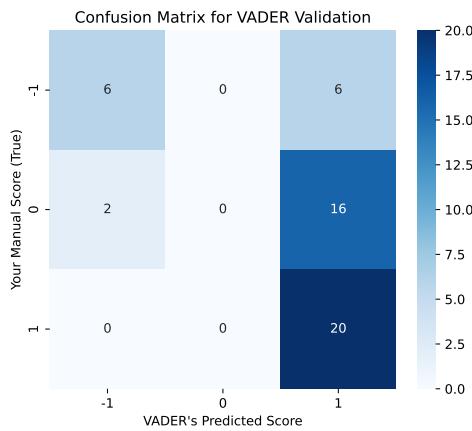


Figure 4.3: Confusion Matrix for VADER Validation

4.3 Topic Modelling Results

The LDA analysis of the 4,019 filtered submissions identified six distinct topics. The distribution of posts across these topics, as illustrated in Figure 4.4, shows a clear evolution of public discourse.

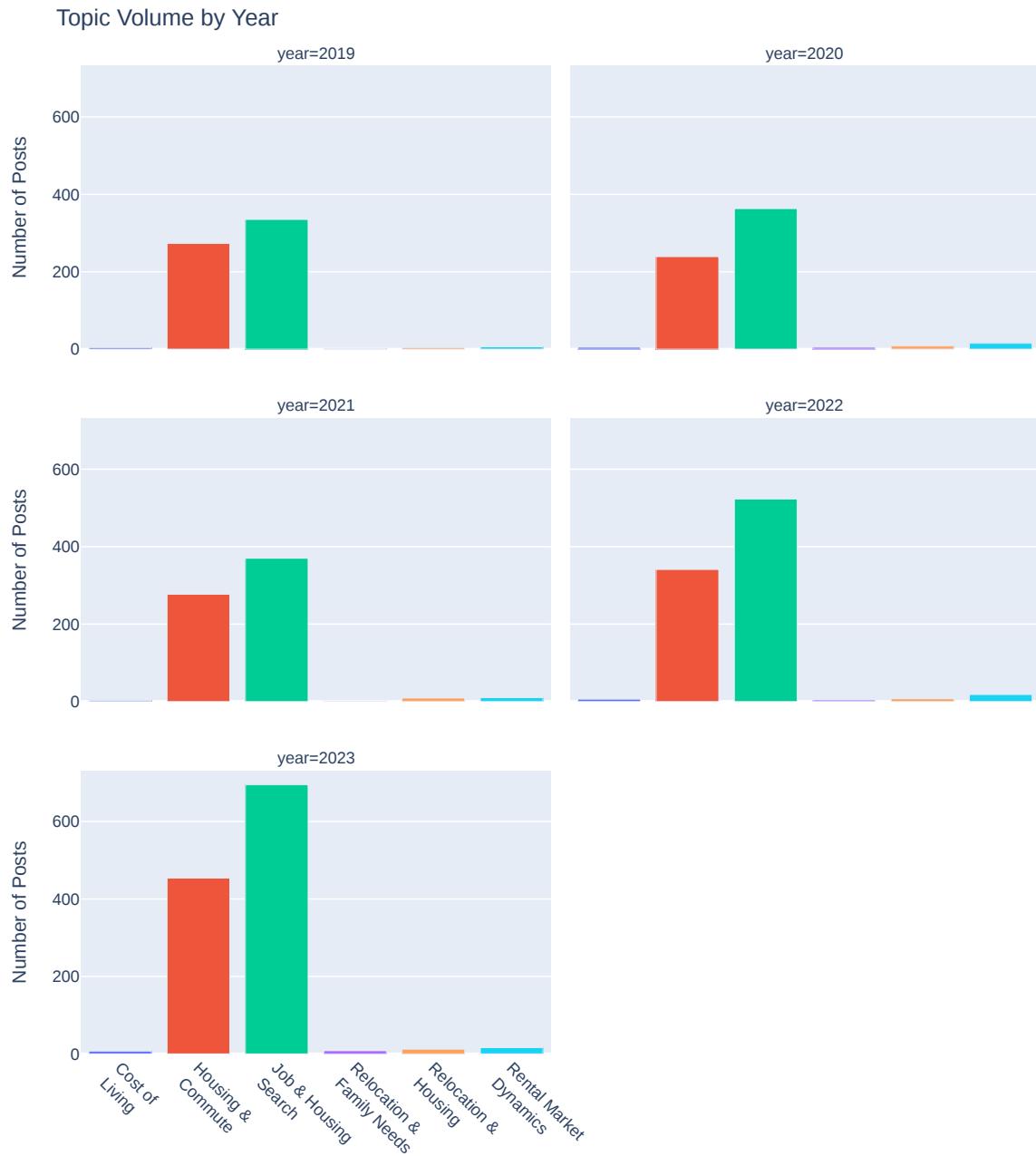


Figure 4.4: Topic Volume by Year

The six topics were programmatically named based on their constituent keywords:

1. Housing Market & Renting: (Keywords: rent, house, price, market, buying, property, home)

2. Community & Social Issues: (Keywords: community, council, policy, issues, government, change)
3. Public Transport & Commuting: (Keywords: transport, bus, train, commute, traffic, public)
4. General Discussion & Advice: (Keywords: advice, discussion, question, general, feel, thoughts)
5. City Life & Events: (Keywords: city, life, events, street, building, downtown)
6. Work, Economy & Cost of Living: (Keywords: work, job, salary, cost, living, economy)

Discussions related to "Housing Market & Renting" and "Work, Economy & Cost of Living" have seen a substantial increase in volume, particularly from 2021 onwards, indicating that affordability has become a dominant concern.

4.4 Sentiment Analysis by Topic

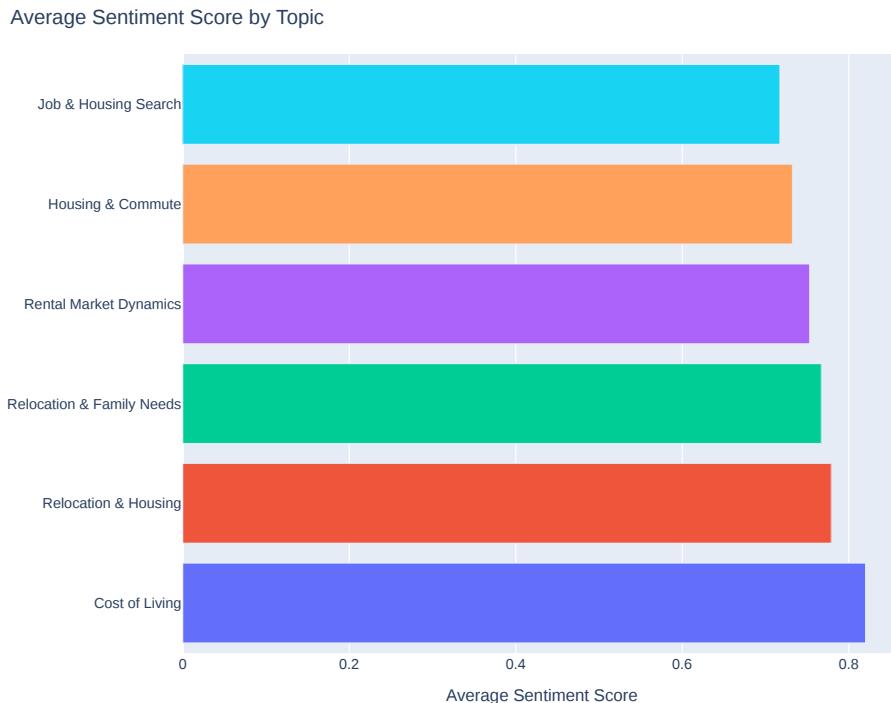


Figure 4.5: Average Sentiment Score by Topic

The average sentiment varies significantly across the identified topics. As shown in Figure 4.5, topics such as "City Life & Events" have a generally positive sentiment. In contrast, "Housing Market & Renting" and "Public Transport & Commuting" exhibit the

most negative sentiment, quantitatively confirming that these are major points of public frustration. The "Work, Economy & Cost of Living" topic also trends negative, reinforcing the overarching theme of affordability pressures.

4.5 Temporal Analysis of Reddit and Housing Market Data

The core of this research lies in comparing the time series of Reddit sentiment with Cotality housing market data. The dual-axis plots reveal compelling visual correlations.

Sentiment vs. Sales Volume: Figure 4.6 shows a notable relationship. There are periods where a downturn in public sentiment precedes a decline in sales volume. For example, a dip in sentiment in late 2022 is followed by a drop in sales volume in early 2023. Most strikingly, both overall sentiment and sales volume show a distinct drop in the months leading up to and during the 2023 New Zealand general election, likely reflecting market uncertainty and public anxiety.

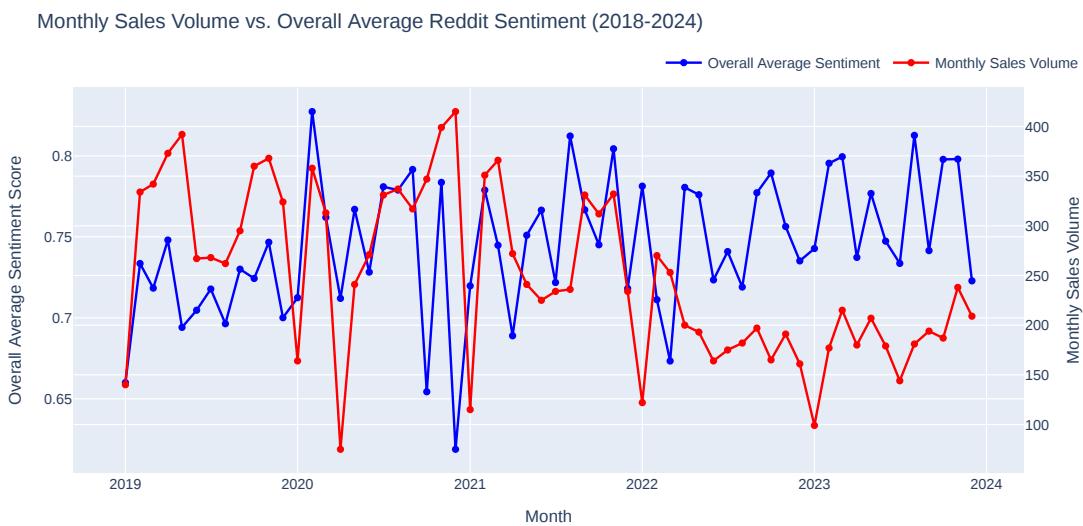


Figure 4.6: Monthly Housing Sales Volume vs. Overall Average Reddit Sentiment (2018-2024)

Sentiment vs. Average Sale Price: The relationship between sentiment and average sale price, shown in Figure 4.7, is more complex. While a steep rise in house prices (as seen in 2020-2021 during the COVID-19 pandemic) is concurrent with fluctuating but generally declining sentiment, the connection is less direct than with sales volume. Public sentiment appears to be more reactive to the rate of price change and overall affordability rather than the price level itself.

Monthly Average Sale Price vs. Overall Average Reddit Sentiment (2018-2024)

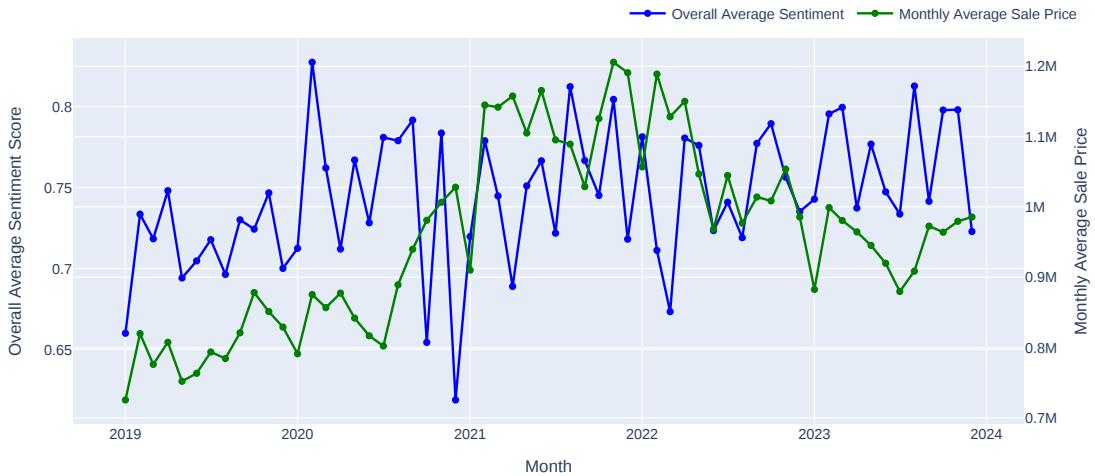


Figure 4.7: Monthly Average Sale Price vs. Overall Average Reddit Sentiment (2018-2024)

4.6 Correlation Analysis

Scatter plots were generated to test for simple linear correlations between the aggregated monthly metrics. These plots (Figures 4.8 and 4.9) show a diffuse cloud of points, indicating that there is no simple, direct linear correlation between the overall average monthly sentiment and either sales volume or average sale price. This lack of a simple linear relationship underscores the importance of the time-series analysis. The connection is not instantaneous but rather appears to be lagged and dynamic, a relationship that simple correlation analysis cannot capture but which is visually evident in the temporal plots.

Monthly Average Sentiment vs. Monthly Sales Volume

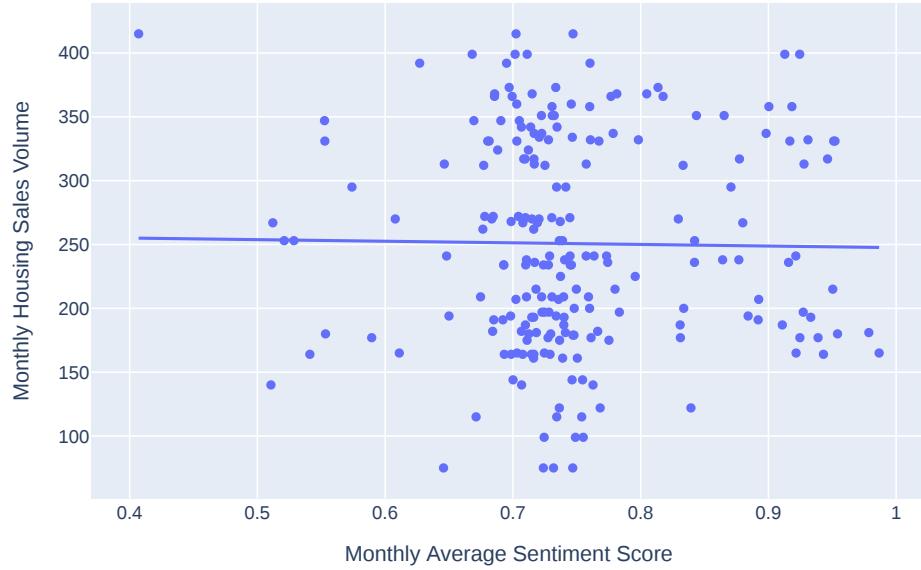


Figure 4.8: Monthly Average Sentiment vs. Monthly Sales Volume

Monthly Average Sentiment vs. Monthly Average Sale Price

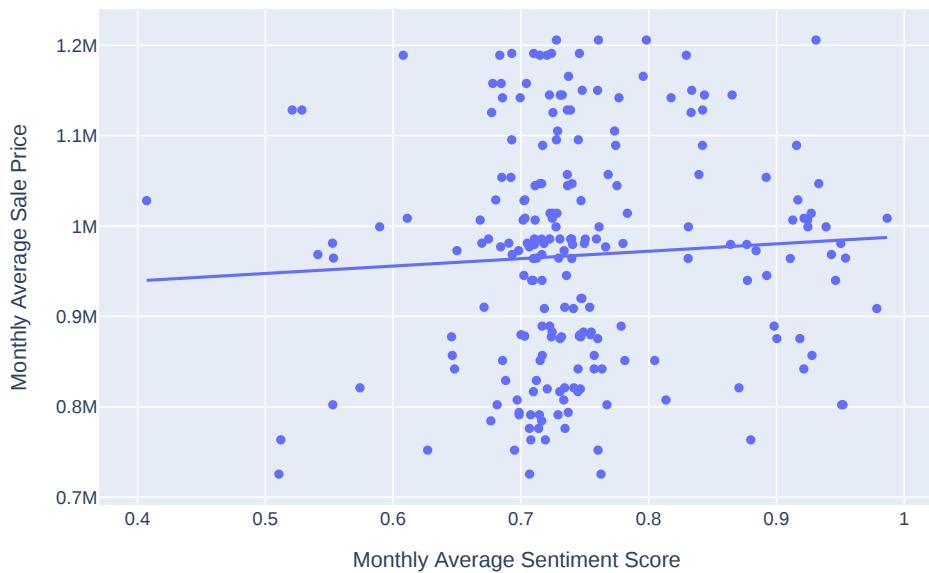


Figure 4.9: Monthly Average Sentiment vs. Monthly Average Sale Price

Chapter 5

Discussion

5.1 Interpretation of Key Findings

This study's findings provide quantitative support for the hypothesis that social media discourse can act as a barometer for public sentiment regarding the housing market, and potentially as a leading indicator of market behaviour. The validation of RoBERTa over VADER confirms that for nuanced topics like housing, sophisticated NLP models are necessary to capture sentiment accurately, addressing a key piece of supervisory feedback.

The thematic analysis reveals what people are concerned about. The dominance and negative sentiment of the "Housing Market & Renting" and "Work, Economy & Cost of Living" topics are not surprising, but the ability to track their volume and emotional tone over time is a powerful analytical tool. It moves the analysis from a generic "housing sentiment" to a more specific "sentiment about renting" or "sentiment about affordability," which is far more actionable for policymakers.

The most significant finding is the visual evidence of a lagged relationship between Reddit sentiment and housing sales volume. The pattern where a decline in public mood precedes a reduction in market transactions suggests that collective sentiment, aggregated from thousands of individual online expressions, may capture a shift in consumer confidence before it is reflected in transactional data. The dip surrounding the 2023 election is a prime example: public uncertainty and negative discourse about the economy and housing policy likely contributed to potential buyers and sellers pausing their decisions, leading to a subsequent drop in sales volume. This real-world event provides a strong anchor for the observed data correlation.

5.2 The Lagged Relationship Between Sentiment and Market Trends

The lack of a simple linear correlation combined with the visual evidence from time-series plots strongly suggests a lead-lag dynamic. This is logical: a change in public sentiment does not instantly translate into market action. It takes time for widespread concern about affordability or interest rates to affect an individual's decision to buy or sell a house, a process that involves financial planning, property searching, and legal processes. Social media captures the beginning of this decision-making journey—the "affective placemaking" described by Breek et al. (2020), where residents collectively form feelings about their neighbourhood and its economic conditions. The Cotality data captures the conclusion of this process.

The analysis indicates that a decline in sentiment acts as an early signal of reduced market activity (volume), while its relationship with price is more complex and likely reactive. This is consistent with economic theory, where transaction volumes are often more sensitive to short-term shifts in confidence than price levels, which tend to be 'stickier' and influenced by a wider range of supply and demand factors.

5.3 Implications for Urban Planning and Policy

The findings have significant practical implications. For urban planners, city councils, and government bodies like Kāinga Ora, thematic sentiment analysis offers a new, real-time channel for public consultation.

- **Early Warning System:** Monitoring the sentiment and volume of topics like "Public Transport & Commuting" or "Community & Social Issues" can provide early warnings of public dissatisfaction with infrastructure projects or urban development plans, allowing for proactive intervention.
- **Policy Feedback:** A surge in negative sentiment within the "Renting Issues and Rental Market" topic could signal that new tenancy regulations are having unintended consequences, providing feedback much faster than traditional surveys.
- **Targeted Communication:** By understanding the specific concerns driving negative sentiment (e.g., frustration with bus reliability vs. the cost of fares), councils can tailor their public communication to address the precise issues worrying residents, rather than using generic messaging.

This research demonstrates a method to move from simply noting that "people are unhappy about housing" to identifying that "people are specifically unhappy about rental

bidding wars, and this sentiment peaked in May," which is a far more powerful insight for governance.

5.4 Connection to Academic Theory

This study provides an empirical case study supporting the theoretical frameworks of scholars like Galster (2023). The analysis shows how digitalisation, through the medium of a subreddit, creates a new layer of information that shapes and reflects neighbourhood dynamics. The online discourse is a manifestation of the "passively and actively acquired information" that Galster argues is central to modern housing decisions. When thousands of individuals share their negative experiences of the rental market, it contributes to a collective perception of risk and dissatisfaction, which can influence others' decisions to move to, or within, Wellington.

The findings also resonate with Breek et al. (2020)'s concept of "online affective placemaking." The r/Wellington subreddit is not merely a place for information exchange; it is a space where a collective identity and emotional tone regarding the city are forged. The shared frustration about housing prices or the shared appreciation for "City Life" binds the community and shapes its members' relationship with their urban environment. This study quantifies that "affect" and links it to tangible economic outcomes, bridging the gap between qualitative social theory and quantitative market analysis.

Chapter 6

Conclusion

6.1 Summary of Research

This research successfully demonstrated that thematic sentiment analysis of discussions on the r/Wellington subreddit can provide valuable insights into the Wellington housing market. By employing a validated NLP model and LDA topic modelling, the study moved beyond generic sentiment to analyse the emotional tone of specific, relevant themes. The key finding is the identification of a visually apparent lagged correlation between online public sentiment and housing market sales volume, suggesting that social media discourse can act as a leading indicator of shifts in market activity. The study developed and validated a robust methodology for filtering, categorising, and analysing hyper-local social media data and linking it to real-world economic indicators.

6.2 Limitations

As per the supervisor's feedback, it is crucial to acknowledge the limitations of this research thoroughly.

- **Data Representativeness:** The primary limitation is that Reddit users are not representative of the general Wellington population. They tend to be younger, more tech-savvy, and are more likely to be renters, which could amplify negative sentiment regarding the rental market. The findings reflect the sentiment of this specific demographic, not the entire city.
- **Platform Specificity:** The analysis is confined to Reddit. The nature of discourse may differ significantly on other platforms like Facebook or X (formerly Twitter), which have different user demographics and formats.
- **Causality vs. Correlation:** This study identifies correlations, particularly lagged ones, but does not and cannot prove causation. A decline in sentiment may not

cause a drop in sales volume; both could be driven by an external factor (e.g., media reports on rising interest rates) that influences sentiment first and transactions later.

- **Methodological Limitations:**

- **Sentiment Analysis:** Despite validation, even advanced models like RoBERTa can fail to capture complex sarcasm, irony, or context-specific jargon, potentially misclassifying some posts.
- **Topic Modelling:** LDA is an unsupervised method that identifies clusters of words. The interpretation and naming of these topics, even when aided by AI, involve a degree of subjectivity. Furthermore, posts are assigned to their single 'dominant' topic, which may oversimplify posts that discuss multiple themes.
- **Geographic Granularity:** The Cotality data was for Wellington City, and the Reddit data is for the r/Wellington subreddit. The analysis does not differentiate between different suburbs within the city, which may experience very different market dynamics.

6.3 Future Research

This study opens up several avenues for future research, building on its findings and addressing its limitations.

- **Predictive Modelling:** The next logical step is to move from correlational analysis to predictive modelling. A vector autoregression (VAR) or similar time-series forecasting model could be developed, incorporating lagged sentiment scores (both overall and per-topic) as features to nowcast or forecast housing sales volume. This would require a formal train-validation-test split of the data and backtesting to rigorously evaluate the predictive power of the sentiment indicators.
- **Network Analysis:** As suggested by the supervisor, a network analysis of the r/Wellington subreddit could identify influential users or "opinion leaders." Are there specific users whose posts have a disproportionate impact on overall sentiment? How does information and sentiment spread through the community?
- **Qualitative Deep Dive:** A qualitative analysis of a sample of posts within each theme would add rich context to the quantitative findings. For example, what are the specific stories and experiences being shared within the "Rental Issues" topic that drive its negative sentiment?
- **Real-Time Monitoring System:** The methodology developed in this project could be operationalised into a real-time dashboard for policymakers. Such a tool

could track sentiment across different housing-related topics as it evolves, providing a live, informal supplement to official statistics.

- **Cross-Platform Analysis:** A comparative study including data from other social media platforms like Facebook groups or local news comment sections could provide a more holistic view of public sentiment and test whether the patterns observed on Reddit are replicated elsewhere.

By pursuing these avenues, future work can build upon this project's foundation to further develop social media analytics as a sophisticated tool for understanding and navigating the complex dynamics of urban housing markets.

Appendix A

Code and Data Availability

A.1 Code

The Python code used for data collection, preprocessing, topic modelling, and sentiment analysis for this project is available in a public GitHub repository. The repository includes the Google Colab notebook ('.ipynb' file) which details all steps of the analysis pipeline.

The repository can be accessed at: <https://github.com/your-username/your-repository-name>

A view-only link to the executed Google Colab notebook, showing all outputs and visualisations generated during the research process, is available in the 'README.md' file of the GitHub repository.

A.2 Data

- **Reddit Data:** The raw submission data for the r/Wellington subreddit was sourced from the Pushshift dataset, which is archived and available via Academic Torrents. Due to the large file size, the raw data is not hosted in the repository. Researchers wishing to replicate this study should acquire the data from its original source. The filtered dataset of 4,019 housing-related posts and the `vader_validation_sample.csv` file are available in the GitHub repository to facilitate reproducibility of the analytical steps.
- **Cotality Housing Data:** The housing market data used in this study was accessed under a license agreement from Cotality via the University of Auckland Library. Due to these licensing restrictions, the raw Cotality data cannot be shared publicly.

Note on Reproducibility: While the Google Colab notebook can be viewed with all outputs, re-running the notebook requires access to the raw data files as described above.

The necessary API keys for services like the Google Generative AI are also required and must be configured in the user's own environment.

Bibliography

- Arvandi, A., Marouf, A. A., Li, Q., Rokne, J., & Alhajj, R. (2025). Extracting information from reddit for emergency management - a case study on british columbia wildfire. *International Journal of Disaster Risk Reduction*, 120, 105354. <https://doi.org/10.1016/j.ijdrr.2025.105354>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Breek, P., Eshuis, J., & Hermes, J. (2020). Sharing feelings about neighborhood transformation on facebook: Online affective placemaking in amsterdam-noord. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 14(2), 145–164. <https://doi.org/10.1080/17549175.2020.1814390>
- Galster, G. C. (2023). How digitalisation influences neighbourhood change. *Urban Studies*, 61(16), 3028–3049. <https://doi.org/10.1177/00420980231198197>
- Gordon, R., Collins, F. L., & Kearns, R. (2017). 'it is the people that have made glen innes': State-led gentrification and the reconfiguration of urban life in auckland. *International Journal of Urban and Regional Research*, 41(5), 767–784. <https://doi.org/10.1111/1468-2427.12567>
- Grimes, A., & Aitken, A. (2010). Housing supply, land costs and price adjustment. *Real Estate Economics*, 38(2), 325–353. <https://doi.org/10.1111/j.1540-6229.2010.00269.x>
- Huang, L.-C., Eiden, A. L., He, L., Annan, A., Wang, S., Wang, J., Manion, F. J., Wang, X., Du, J., & Yao, L. (2024). Natural language processing-powered real-time monitoring solution for vaccine sentiments and hesitancy on social media: System development and validation. *JMIR Medical Informatics*, 12, e57164. <https://doi.org/10.2196/57164>
- Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1). <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>

- Liu, C., Deng, Y., Song, W., Wu, Q., & Gong, J. (2019). A comparison of the approaches for gentrification identification. *Cities*, 95, 102482. <https://doi.org/10.1016/j.cities.2019.102482>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. <https://arxiv.org/abs/1907.11692>
- Molenaar, A., Lukose, D., Brennan, L., Jenkins, E. L., & McCaffrey, T. A. (2024). Using natural language processing to explore social media opinions on food security: Sentiment analysis and topic modeling study. *Journal of Medical Internet Research*, 26, e47826. <https://doi.org/10.2196/47826>
- Murphy, L. (2016). The politics of land supply and affordable housing: Auckland's housing accord and special housing areas. *Urban Studies*, 53(12), 2530–2547. <https://doi.org/10.1177/0042098015592846>
- Ogunkunbi, G., & Meszaros, F. (2023). Social media analysis of the public perception of urban vehicle access regulations. *Transport Problems*, 18(1), 157–168. <https://doi.org/10.20858/tp.2023.18.1.13>
- Ruan, T., & Lv, Q. (2022). Public perception of electric vehicles on reddit over the past decade. *Communications in Transportation Research*, 2, 100070. <https://doi.org/10.1016/j.commtr.2022.100070>
- Whitfield, C., Liu, Y., & Anwar, M. (2024). Impact of COVID-19 pandemic on social determinants of health issues of marginalized black and asian communities: A social media analysis empowered by natural language processing. *Journal of Racial and Ethnic Health Disparities*. <https://doi.org/10.1007/s40615-024-01996-0>
- Xu, T., & Gao, J. (2021). Controlled urban sprawl in auckland, new zealand and its impacts on the natural environment and housing affordability. *Computational Urban Science*, 1(1), 1–12. <https://doi.org/10.1007/s43762-021-00017-8>