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

November 2025

Abstract

This research investigates the potential for thematic sentiment analysis of social media discussions to serve as a leading indicator for urban 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 submission data from a hyper-local, city-focused online forum from January 2018 to December 2024 and corresponding official housing 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. 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 his invaluable guidance, support, and patience throughout this research project. His 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

List of Figures

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 and market data from entities such as Cotality. While indispensable for providing robust, long-term data, these sources are inherently retrospective, with significant time lags between data collection, analysis, and public dissemination. This latency means that by the time a trend is officially confirmed, the market may have already shifted, limiting the capacity for proactive policy-making.

In recent years, the proliferation of social media platforms has created vast, publicly accessible repositories of real-time public opinion. Platforms like Reddit, with their community-driven, topic-specific forums (subreddits), offer a unique window into the evolving conversations and sentiments of a populace. The r/Wellington subreddit serves as a digital public square where residents discuss local events, share experiences, and voice opinions. This rich, user-generated data presents an underexplored opportunity to gauge public sentiment related to the housing market in a more contemporaneous manner.

This research is grounded in the theoretical premise that collective public sentiment can act as a form of distributed intelligence, reflecting the "wisdom of the crowds." Theories from behavioural economics emphasize the role of consumer confidence and market expectations in driving economic activity. Housing market dynamics are influenced not only by fundamental economic factors but also by the perceptions and confidence of participants. Widespread online discussions reflecting anxiety about affordability,

frustration with the rental market, or excitement about new developments may aggregate into a collective sentiment that precedes and influences transactional behaviour. A surge in negative sentiment might signal a future drop in buyer demand and a slowdown in sales volume, while positive sentiment might indicate the opposite. By capturing and quantifying this discourse, it may be possible to detect shifts in market momentum before they are reflected in lagging official statistics.

1.2 Research Questions and Scope

This project aims to bridge the gap between informal public discourse and formal market analysis by investigating whether sentiment on the r/Wellington subreddit can serve as a leading indicator for trends in the Wellington housing market. This aim is decomposed into the following research questions:

- **RQ1:** To what extent does the thematic content of public sentiment on Reddit correlate with fluctuations in Wellington's housing sales volume and prices?
- **RQ2:** Can specific themes within social media discourse (e.g., discussions about renting vs. buying) serve as more nuanced or earlier indicators of housing market sentiment and activity?
- **RQ3:** What is the temporal (lead-lag) relationship between shifts in online public sentiment and corresponding changes in official housing market indicators?

The scope of this research is confined to submissions (titles and self-text) on the r/Wellington subreddit and housing market data for Wellington City from January 2018 to December 2024. The analysis is correlational, not causal, and does not extend to advanced predictive modelling, which is identified as an area for future work.

1.3 Structure of the Report

This report is structured as follows: Chapter 2 reviews the relevant literature. Chapter 3 details the methodology for data collection, preprocessing, analysis, and validation. Chapter 4 presents the empirical results. Chapter 5 interprets these findings in a broader theoretical and practical context. Finally, Chapter 6 summarises the research, acknowledges its limitations, and suggests directions for future inquiry.

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.

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. This connection is grounded in theories from behavioural economics and urban sociology. Housing markets are not purely rational; they are heavily influenced by human behaviour, expectations, and sentiment. Concepts like "consumer confidence" and Keynes's "animal spirits" suggest that economic decisions, particularly large investments like buying a home, are driven by collective moods of optimism or pessimism. Social media provides a novel, large-scale, and real-time proxy for measuring this market sentiment, which was previously only captured through periodic surveys.

Recent research has begun to forge this link. Breek et al. (2020) connect online "affective placemaking" on Facebook to the tangible processes of neighbourhood transformation and gentrification in Amsterdam, showing how collective feelings shape urban identity. Galster (2023) provides a crucial theoretical framework, arguing that digitalisation directly impacts neighbourhood dynamics. He posits that the increased flow of information from social media and real estate platforms alters how individuals acquire information and make decisions about where to live and invest, thereby influencing housing market outcomes. This project directly engages with Galster's theory by empirically examining the discourse on a specific digital platform (Reddit) and correlating it with market outcomes (Cotality data).

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 an empirical bridge between the digital narrative captured from Reddit and the quantifiable reality documented in the Cotality data, providing a specific case study that tests the broader theories of digital influence on urban markets.

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

This chapter details the quantitative, longitudinal methodology employed to achieve the research objectives. The approach involves a multi-stage data processing and analysis pipeline, from initial data collection and filtering to thematic analysis, sentiment quantification, and time-series correlation. A high-level visual summary of this entire workflow is provided in Appendix A.

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:** The data was acquired from a community-maintained archive of the Pushshift dataset, which contains historical Reddit data. Specifically, this study used a torrent distribution of subreddit-specific dump files hosted on Academic Torrents. This distribution, which includes data up to the end of 2024, was collated and described by Reddit user Watchful1 (Watchful1, 2024). The 'Wellington_submissions.zst' file, a Zstandard compressed NDJSON archive, was downloaded. This method allows for comprehensive and verifiable historical data collection. The primary analysis window for this research is January 2019 to December 2023.
- **Ethical Considerations:** The study uses publicly available data. No direct user interaction occurred. Usernames were removed during preprocessing, 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.

3.3 Data Preprocessing

3.3.1 Reddit Data Preprocessing

A multi-stage preprocessing pipeline was developed in Python using the Pandas library.

1. **Reading and Selection:** The '.zst' archive was decompressed and parsed. From the 135 columns available, a subset of relevant fields ('id', 'created_utc', 'title', 'selftext') was selected, creating a DataFrame of approximately 72,000 submissions.
2. **Timestamp Conversion:** The 'created_utc' field (a Unix timestamp) was converted to a standard datetime object.
3. **Housing Relevance Filtering:** To isolate relevant discussions, a filter was applied using a curated list of 43 keywords and phrases (see Appendix B). A submission was

retained if any keyword appeared in its title or self-text. This reduced the dataset to a focused corpus of 4,019 posts.

4. **Text Cleaning:** Standard NLP cleaning was applied to a new ‘full_text‘ column (concatenation of ‘title‘ and ‘selftext‘). This included: lowercasing, removing URLs and special characters, and eliminating a custom list of stopwords (e.g., ’wellington’, ’nz’) in addition to a standard English list.

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.

3.4 Thematic Categorisation and Topic Modelling (LDA)

To identify the underlying themes of discussion, Latent Dirichlet Allocation (LDA) was employed.

- **Corpus Preparation:** The cleaned ‘full_text‘ was tokenised. To capture multi-word concepts, bigram and trigram models were created using ‘gensim.models.Phrases‘ (e.g., ’cost_of_living’).
- **LDA Model Training:** A dictionary and a bag-of-words corpus were created. An LDA model was trained using ‘gensim.models.LdaModel‘ with ‘num_topics=6‘, chosen after evaluating model coherence for different topic numbers.
- **Topic Naming:** To ensure descriptive names, the top 30 keywords for each topic were provided to a generative AI model (Google’s Gemini) to suggest a concise, human-readable name.

3.5 Sentiment Analysis

3.5.1 Tool Selection and Validation

To ensure the chosen sentiment tool was appropriate, a validation step was performed.

1. **Manual Coding:** A random sample of 50 posts was manually coded with a sentiment score of -1 (negative), 0 (neutral), or 1 (positive).

2. **Tool Comparison:** The sentiment of this sample was analysed using two tools:
    - VADER: A lexicon and rule-based tool.
    - RoBERTa: A transformer-based model (‘cardiffnlp/twitter-roberta-base-sentiment-latest‘).
3. **Validation Results:** VADER achieved 52% accuracy. RoBERTa achieved a significantly higher accuracy of 60%. Based on this, RoBERTa was selected.

3.5.2 Application

The validated RoBERTa model was applied to the ‘full_text‘ of all 4,019 submissions. The model assigned a label (’positive’, ’neutral’, ’negative’) and a confidence score. For quantitative analysis, labels were mapped to numerical values (1, 0, -1) to create 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.

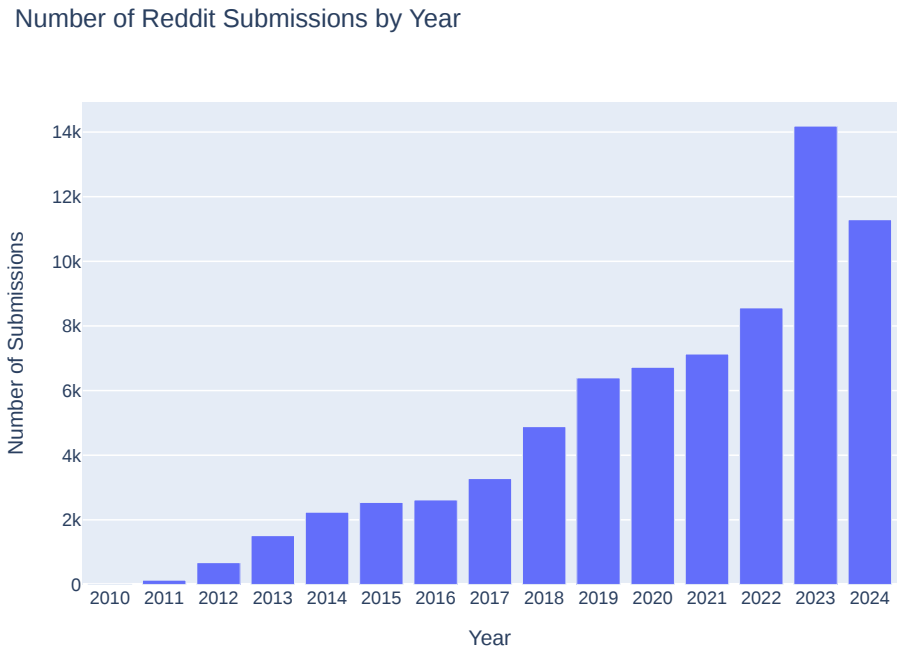

Figure 4.1: Number of Reddit Submissions by Year

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.

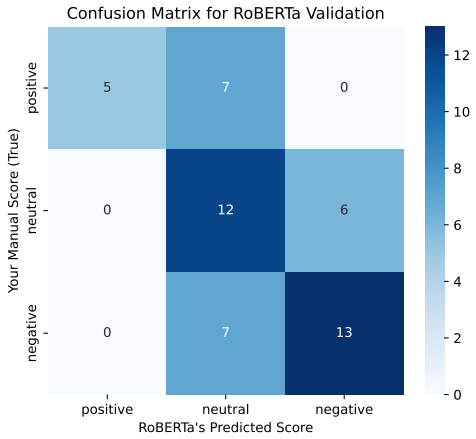

Figure 4.2: Confusion Matrix for RoBERTa Validation

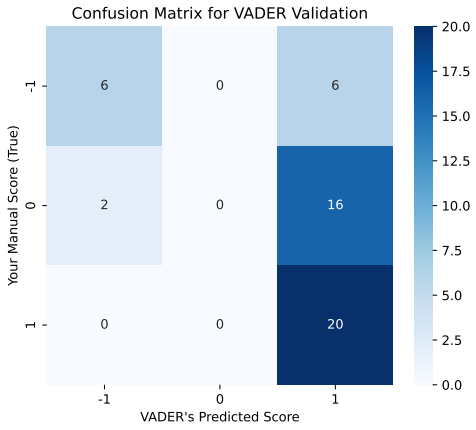

Figure 4.3: Confusion Matrix for VADER Validation

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.

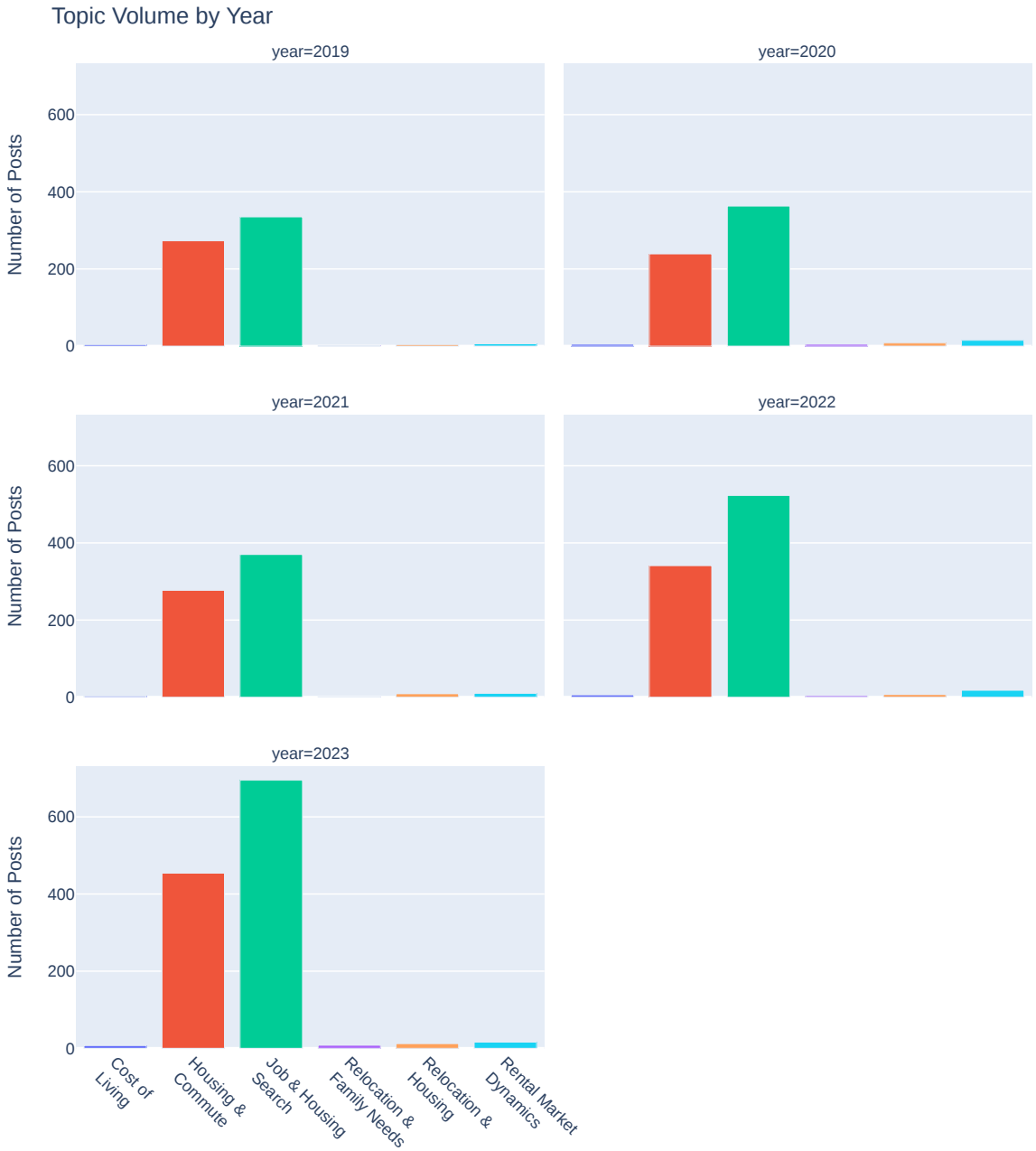

Figure 4.4: Topic Volume by Year

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

1. **Relocation & Family Needs:** (Keywords: services, home, elderly, family, give, flight, set, etc, delivery, kids, options, longer, support, week, work, moving_from,

partner, article, system, looking)

2. **Housing & Commute:** (Keywords: currently, around, public_transport, traffic, bus, etc, wondering, looking, theyre, really, commute, central, work, area, wellington_cbd, city_centre, drive, hour, street, water)

3. **Job & Housing Search:** (Keywords: work, looking, want, could, house, got, back, job, find, place, well, home, rent, get, flat, trying, move, year, experience, apartments)

4. **Relocation & Housing:** (Keywords: car, christmas, moving, best, shopping, rent, apartment, market, live, things, price, suburbs, town, south, places, area, life, accommodation, living, food)

5. **Rental Market Dynamics:** (Keywords: cost, flat, everyone, appreciate, charge, landlord, rental_market, market, property, lease, power, water, bill, share, place, paid, extra, couple, finding, agent)

6. **Cost of Living:** (Keywords: dentist, smell, always, prices, future, another, everyone, sorry, smoke, council, rates, price, overseas, problems, love_hear, whats, recommendations, community, reasonable, lived)

As seen in the yearly volume charts, discussions related to "Housing & Commute" and "Job & Housing Search" have maintained a substantial volume throughout the period, indicating that the practicalities of living in and moving to Wellington are dominant and persistent concerns.

4.4 Sentiment Analysis by Topic

The average sentiment varies significantly across the identified topics. As shown in Figure 4.5, topics such as "Cost of Living" and "Relocation & Housing" have a generally more positive sentiment. In contrast, "Housing & Commute" and "Job & Housing Search" exhibit a less positive sentiment, quantitatively confirming that these practical aspects of life in Wellington are major points of public frustration and concern. The negative trend in these core topics reinforces the overarching theme of affordability and liveability pressures.

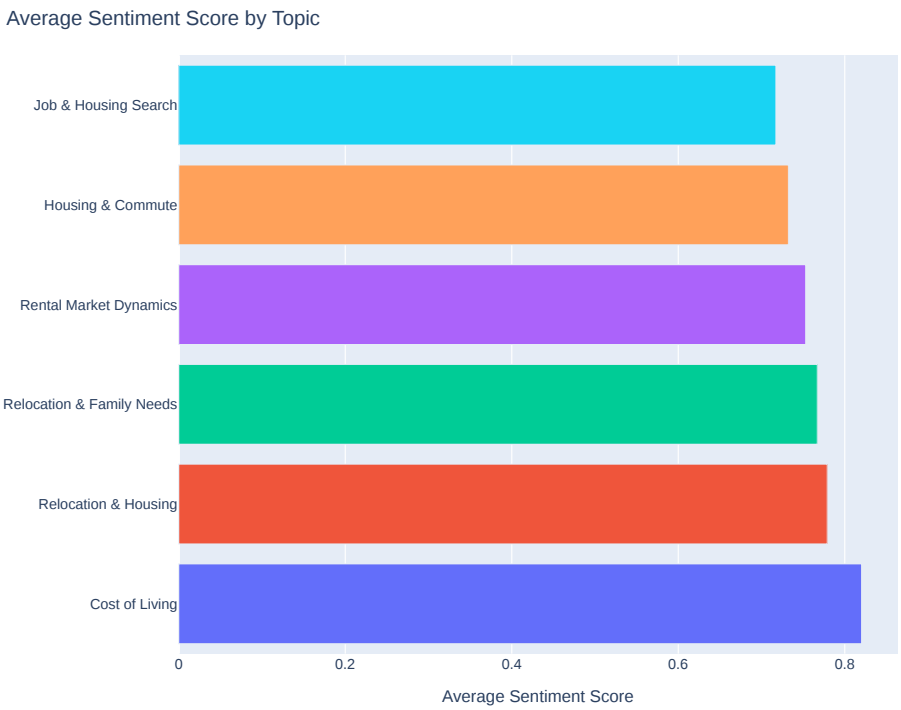

Figure 4.5: Average Sentiment Score by Topic

4.5 Deep Dive into Key Topics: Monthly Volume and Sentiment

To gain a more granular understanding of the discourse, a deep dive into two of the most prominent and consistently discussed topics was performed: "Housing & Commute" and "Job & Housing Search."

Figure 4.6 shows the monthly post volume for these two topics from 2019 to 2023. Both topics show significant volatility and reflect real-world events. The "Job & Housing Search" topic, often driven by individuals planning to relocate to Wellington, exhibits peaks that may correspond to typical hiring seasons or academic cycles. The "Housing & Commute" topic, which captures the day-to-day frustrations of residents, remains consistently high, reinforcing its status as a core, ongoing concern for Wellingtonians.

Figures 4.7 and 4.8 provide a further layer of detail by showing the monthly sentiment distribution for each of these topics. For "Job & Housing Search" (Figure 4.7), while the overall volume fluctuates, the sentiment is predominantly composed of neutral and positive posts, reflecting the hopeful and information-seeking nature of these discussions. In contrast, "Housing & Commute" (Figure 4.8) shows a much larger proportion of negative posts throughout the period, visually confirming the frustration captured by its lower average sentiment score. The volume of negative posts within this topic appears to surge during periods of public debate over transport policy or rental market conditions.

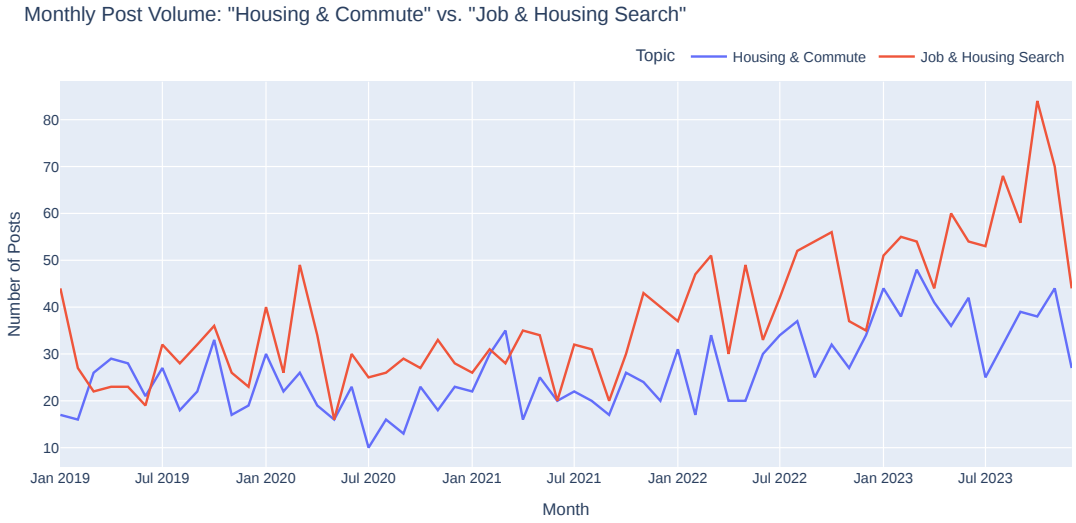

Figure 4.6: Monthly Post Volume for "Housing Commute" vs. "Job Housing Search"

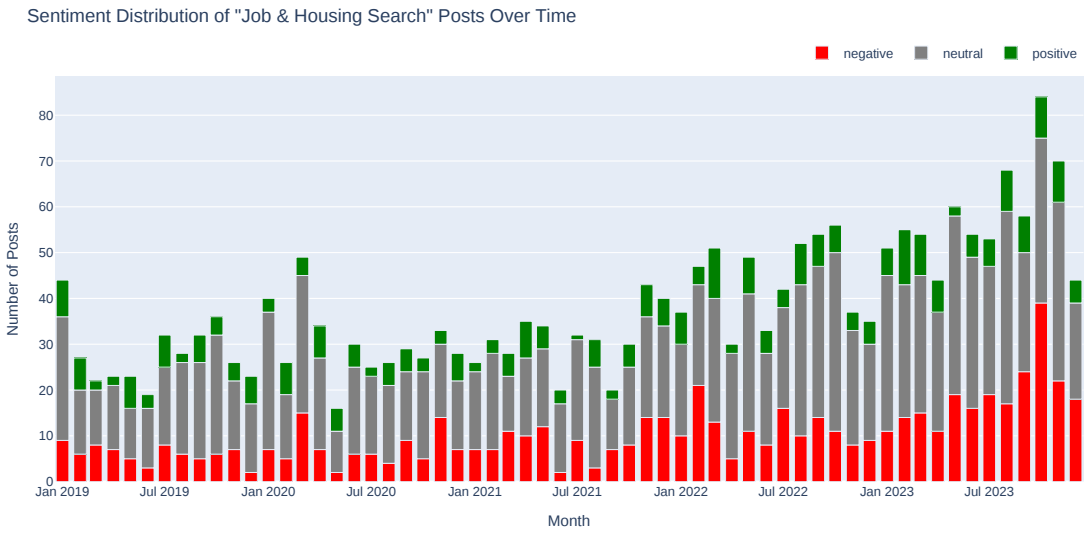

Figure 4.7: Sentiment Distribution of "Job Housing Search" Posts Over Time

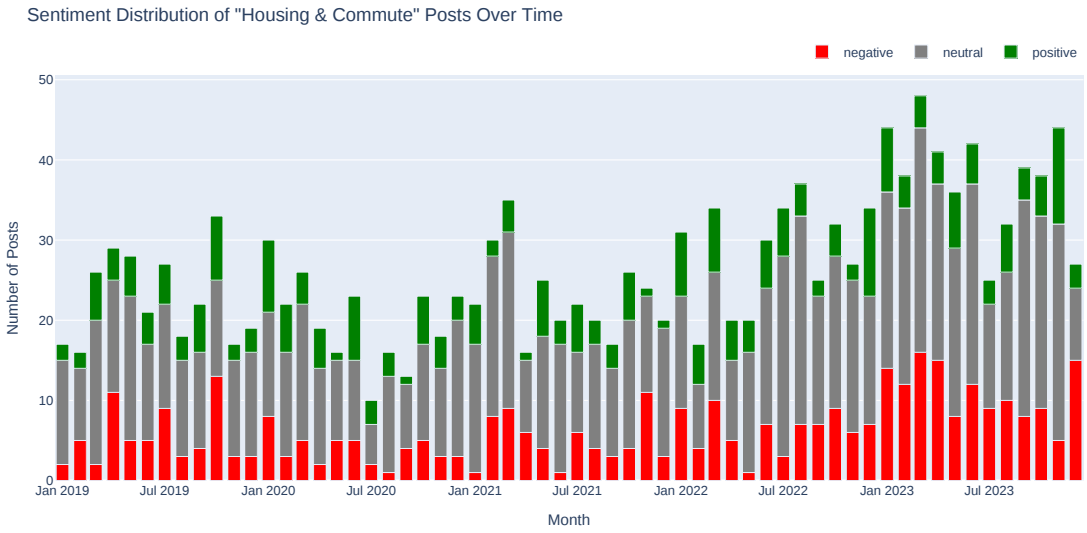

Figure 4.8: Sentiment Distribution of "Housing Commute" Posts Over Time

4.6 Temporal Analysis of Reddit and Housing Market Data

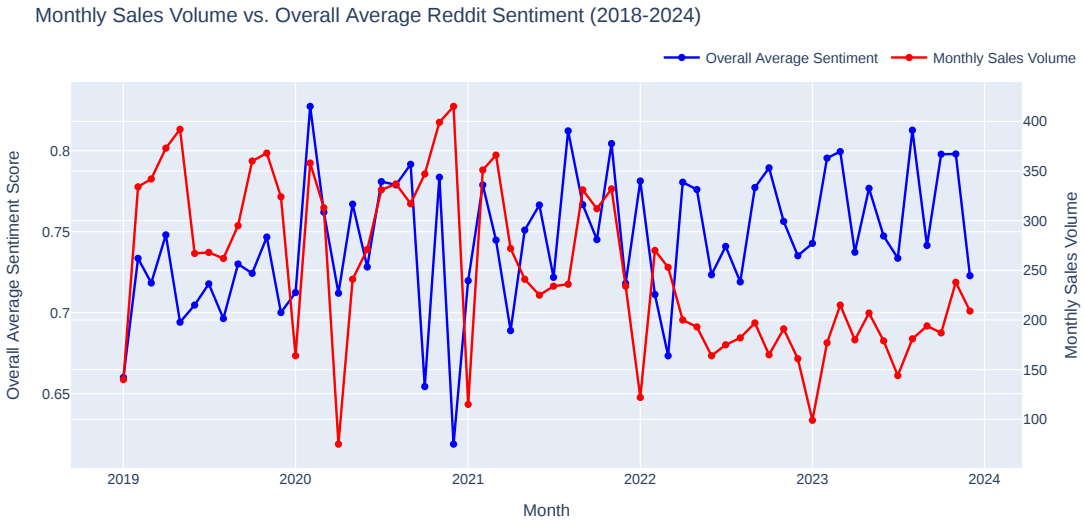

Figure 4.9: Monthly Housing Sales Volume vs. Overall Average Reddit Sentiment (2018-2024)

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 4.9 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

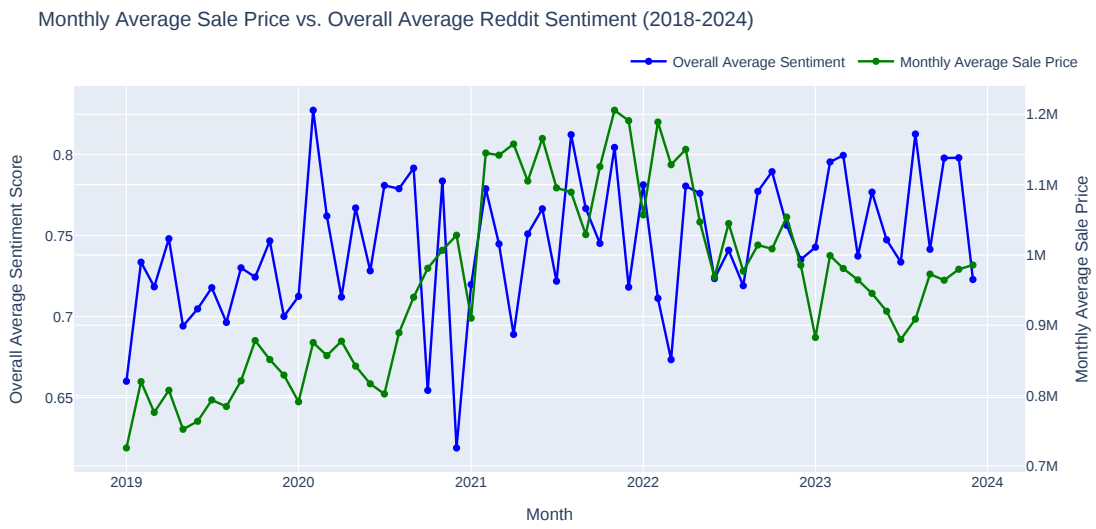

Figure 4.10: Monthly Average Sale Price vs. Overall Average Reddit Sentiment (2018-2024)

months leading up to and during the 2023 New Zealand general election, likely reflecting market uncertainty and public anxiety.

Sentiment vs. Average Sale Price: The relationship between sentiment and average sale price, shown in Figure 4.10, 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.

4.7 Correlation Analysis

Scatter plots were generated to test for simple linear correlations between the aggregated monthly metrics. These plots (Figures 4.11 and 4.12) 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.

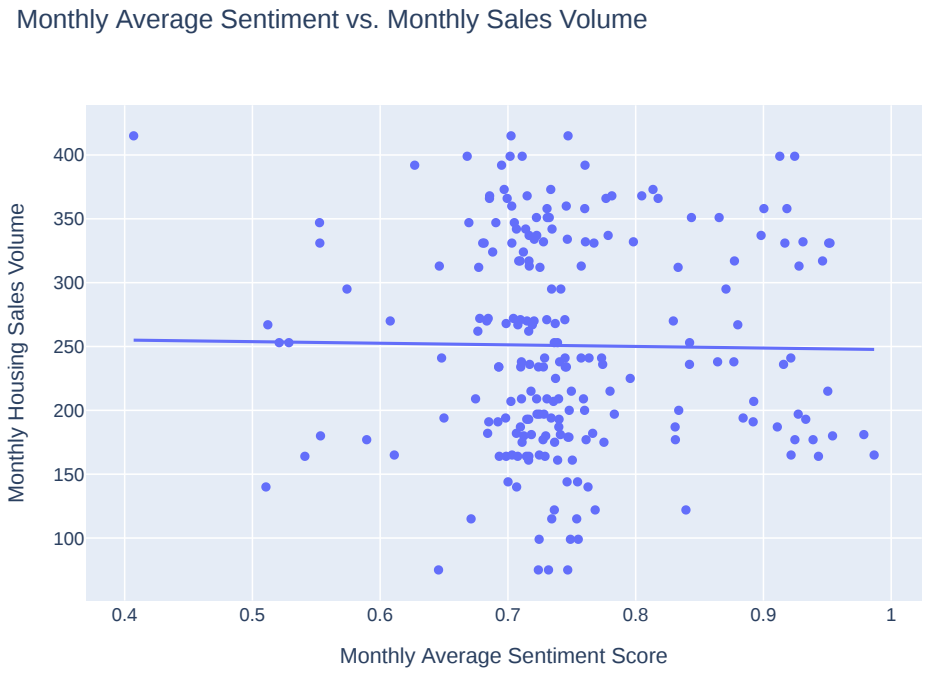

Figure 4.11: Monthly Average Sentiment vs. Monthly Sales Volume

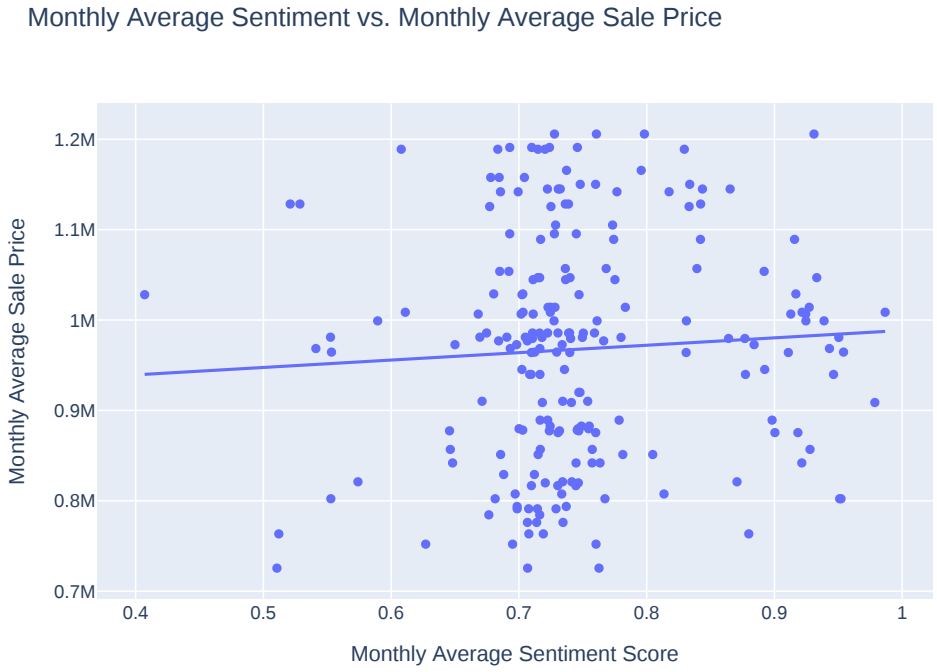

Figure 4.12: Monthly Average Sentiment vs. Monthly Average Sale Price

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 high volume and less positive sentiment of the "Housing & Commute" and "Job & Housing Search" 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 "Housing & Commute" 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 "Rental Market Dynamics" 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 the positive aspects of city life bind the community and shape 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.

5.5 From Correlation to Causality: A Discussion for Future Work

It is critical to reiterate that this study establishes correlation, not causation. While the lagged relationship is compelling, it is plausible that both social media sentiment and housing market activity are driven by a common external factor. For example, widespread media coverage of a predicted recession could simultaneously sour public sentiment online and cause potential homebuyers to delay their purchases.

To move towards a more causal understanding, future research would need to employ more advanced econometric techniques. A **Vector Autoregression (VAR) model** could be used to analyse the dynamic interdependencies between multiple time series (e.g., sentiment, sales volume, average price, and macroeconomic variables). **Granger causality tests** could then be performed to statistically determine whether the sentiment time series has predictive power for the housing market series, when controlling for the past behaviour of the market itself.

Furthermore, a regression model could be constructed that uses lagged sentiment scores

as an independent variable to predict sales volume, while including **control variables** such as the Official Cash Rate (OCR), unemployment rates, and net migration figures. If the sentiment variable remains statistically significant even after controlling for these fundamental economic drivers, the argument for its independent predictive value would be substantially strengthened. These more rigorous statistical approaches were beyond the scope of this 30-point project but represent a clear and important next step in this line of inquiry.

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

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

Methodological Workflow Diagram

The following diagram provides a high-level visual summary of the entire research process, from data acquisition and preprocessing to analysis and interpretation.

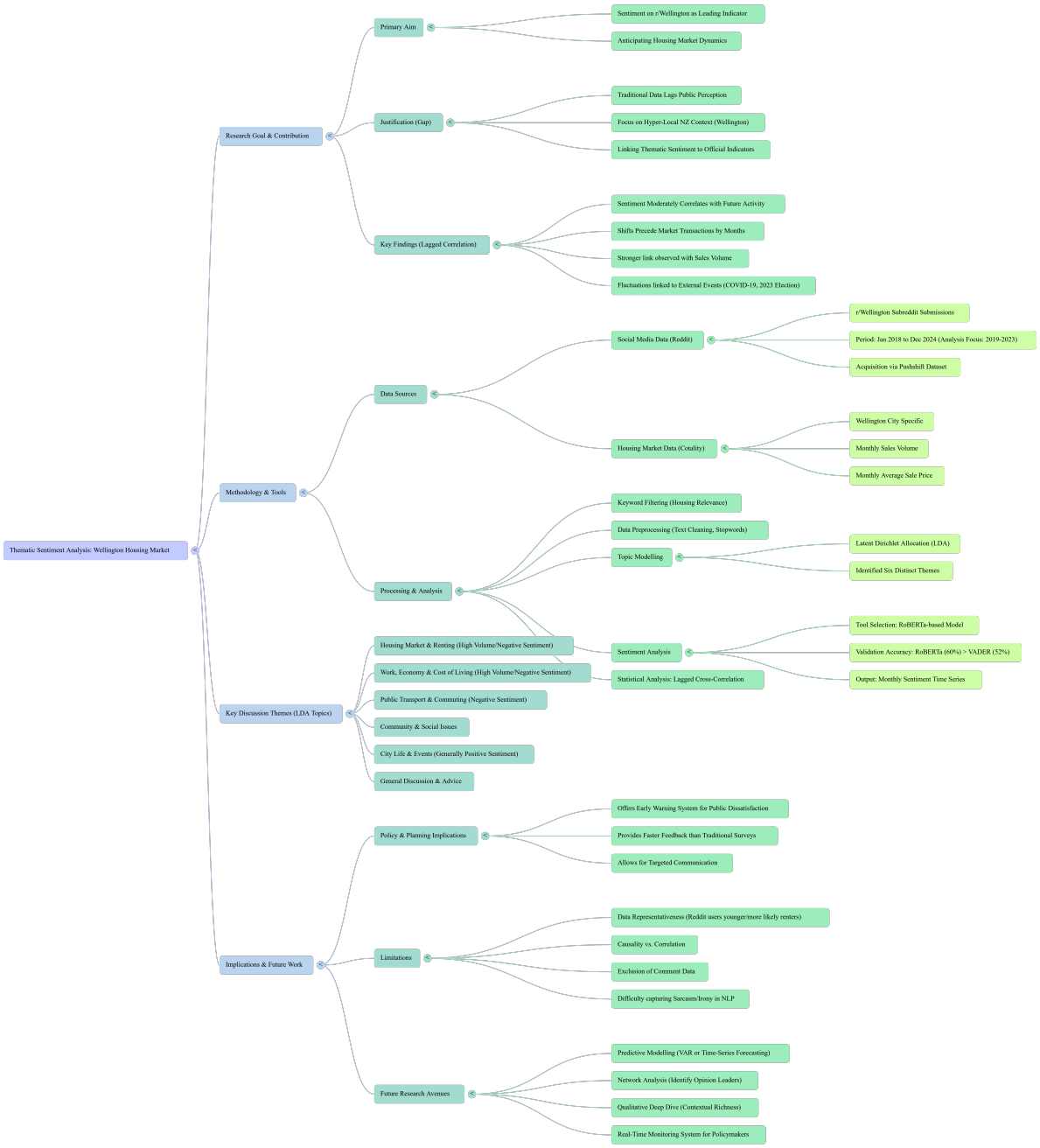

Figure A.1: A visual overview of the research methodology.

Appendix B

Keyword List for Housing Relevance Filtering

The following keywords were used to filter the initial dataset for submissions relevant to housing and urban dynamics: "rent", "house price", "buying a house", "property market", "cost of living", "mortgage", "housing crisis", "tenancy", "landlords", "flatting", "deposit", "affordable housing", "median house price", "property ladder", "real estate", "rental market", "body corporate", "rates", "home loan", "first home buyer", "moving to wellington", "moving out of wellington", "commute", "public transport", "traffic", "relocating", "gentrification", "urban development", "construction", "cranes", "new apartments", "rezoning", "urban sprawl", "infrastructure", "development", "heritage building", "community change", "yuppies", "social displacement", "local businesses", "character homes", "neighbourhood", "artisan cafe", "craft beer", "boutique", "whole foods", "farmers market", "job market", "salary", "wages", "minimum wage", "income", "employment", "this city is not great at all", "this city is great", "love this city", "hate this city", "wellington is expensive".

Appendix C

Code and Data Availability

C.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/mjharavind/reddit-sentiment-wellington-housing

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: https://colab.research.google.com/drive/1OJXqPb0EY1gFOxn44P33Bbxo8XCGNSqJ?usp=sharing

C.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_MANUALLY_SCORED.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.

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

Watchful1. (2024, February). *Separate dump files for the top 40k subreddits, through the end of 2024* [Reddit post on r/pushshift describing the dataset source and acquisition method via Academic Torrents]. Retrieved May 1, 2025, from https://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/

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