Transforming Urban Planning through Machine Learning: A Study on Planning Application Classification using Natural Language Processing

Yang Lin, William Thackway*, Balamurugan Soundararaj, Serryn Eagleson, Hoon Han, Christopher Pettit

***w.thackway@unsw.edu.au**

October 2024

Abstract: Planning for sustainable urban growth is a pressing challenge facing many cities. Investigating proposed changes to the built environment can provide planners and policymakers information to understand future urban development trends and related infrastructure requirements. It is in this context we have developed a novel urban analytics approach that utilises planning applications (PAs) data and Natural Language Processing (NLP) techniques to forecast the housing supply pipeline in Australia. Firstly, we implement a data processing pipeline which scrapes, geocodes, and filters PA data from council websites and planning portals to provide the first nationally available daily dataset of PAs that are currently under consideration. Secondly, we classify the collected PAs into four distinct urban development categories, selected based on infrastructure planning provisioning requirements. Of the five model architectures tested, we found that the fine-tuned DeBERTA-v3 model achieves the best performance with an accuracy and F1-score of 0.944. This demonstrates the suitability of fine-tuned Pre-trained Language Models (PLMs) for planning text classification tasks. Finally, the model is applied to classify and map urban development trends in Australia's two largest cities, Sydney and Melbourne, from 2021-2022 and 2023-2024. The mapping affirms a face-validation test of the classification model and demonstrates the utility of PA insights for planners. Holistically, the paper demonstrates the potential for NLP to enrich urban analytics through the integration of previously inaccessible planning text data into planning analysis and decisions.

***Keywords**: urban development; planning application; natural language processing; deep learning; pre-learned model; housing supply*

1. Introduction

Housing stress is a global problem, predicted by the World Bank to impact 1.6 billion people by 2025 (Caliyurt, 2022). In Australia, as in many other western countries such as the United States and United Kingdom, the financialization of housing, increases in net immigration, and unfavourable macroeconomic conditions have led to housing supply shortages (Phibbs & Gurran, 2021). Moreover, the provision of services and infrastructure to support new housing supply is often insufficient (Kellet & Nunnington, 2019). In recent years, the government has had a diminishing role in financing and building basic residential infrastructure such as roads, schools and hospitals, leaving greater responsibility for developers to design and manage these services in growth areas (Kellet & Nunnington, 2019). Resultingly, social and transport services are consistently undersupplied in new build Australian suburbs compared to established areas, leading to poor resident experiences (Gunn et al., 2020). The cumulative result of these concurrent factors has been inadequate supply of new housing and residential infrastructure in urban growth areas.

It is in this context that understanding the housing supply pipeline can provide valuable insights to planners around where and when new developments are occurring, and the infrastructure requirements for different types of urban developments. In this paper, we propose analysing Planning Applications (PAs) to understand Australia's housing supply pipeline. PAs are an application process required for landowners to build new developments, subdivide land, or change the land use of a property (Murray, 2023). These applications are publicly available on planning portals and council websites, providing rich and freely accessible information on application dates, development descriptions, and estimated costs of potential developments. Similar application processes exist under different names in other international contexts, such as "building permits" and "permitted development rights" in the United States and United Kingdom respectively (Fauth & Soibelman, 2022; Ferm et al., 2024). We propose an urban analytics approach to classify PAs into four distinct urban development categories (commercial development, new housing development, subdivision and land development, alterations and modifications) to help planners understand the types and distribution of property developments. These categories were selected from an infrastructure provisioning perspective to understand the planning implications and needs of different urban development types (Uddin et al., 2022).

Analysing PAs on a metropolitan scale can offer two major insights for policymakers and planners. Firstly, PAs can be used as a measure of new housing supply in a city. For example, Phibbs & Gurran (2021) investigated permitting intensity (measured as the number of new dwellings permitted through PAs compared to existing housing stock) to analyse new housing developments in Sydney over time. This urban analytics approach has also been applied to understand the responsiveness of housing markets to policy incentives and market changes in US cities such as Atlanta and Houston (Glaeser &

Gyouro, 2018). Given the lag between planning applications, planning approvals, and construction, PAs offer the potential to forecast housing supply pipelines over the next 1-2 years (Phibbs & Gurran, 2021). This can help governments understand the responsiveness of private investors and commercial developers to housing policy incentives, and to guide decisions about future housing development policy.

Secondly, understanding different PA development types can help guide decisions around infrastructure requirements in growth areas. For example, in areas with increased commercial development applications, planners may consider improving pedestrian traffic flow, promoting retail and hospitality businesses, and maintaining activity during outside-work hours (Wang & Niu, 2019). In areas with dense new dwelling applications, Uddin et al. (2022) stress the importance of providing job opportunities, health services, transport connectivity, and sufficient educational and cultural facilities. Similar considerations apply for areas with land development and subdivision applications, although the greater likelihood of these approvals occurring in lower density urban fringe areas means the range of transport and service provision options considered may vary (Shahzad et al., 2022). Finally, while alterations and modifications may not directly imply an increase or change in the population mix of an area, dense clustering of modifications may imply socioeconomic change and are nonetheless a useful trend to monitor (Thackway et al., 2023a). Hence, analysing PAs can provide both high-level insights on future housing supply and more targeted insights to inform infrastructure planning at a local level.

While there is a growing body of work internationally investigating the impact of PAs on housing supply and housing prices (e.g., Gabriel & Kung, 2024; van der Kooi, 2024), attempts to aggregate PA data in Australia have been limited to a handful of more recent studies. Zhu et al. (2022) and Murray (2023) analysed structured government PA datasets for Sydney and Queensland respectively, identifying trends in the supply and timeliness of PAs. Both studies offered insights into the spatial and temporal trends of housing development, demonstrating the potential for PAs to enhance planners' understanding of the housing pipeline. However, the studies were limited to the date, cost and location fields contained within the structured csv datasets. Resultingly, the studies analysed overall PA trends, however, did not distinguish between different types of urban development. Contrastingly, unstructured text data contained within PA descriptions may capture detailed property attributes needed to categorise urban development types, such as land use, building form, or construction notes. Accessing these unstructured data requires researchers to interact directly with PA data published on local council websites and government portals.

However, direct interaction with unstructured PA data has historically been limited by the difficulty in processing and standardising PA data sources. In Australia, PAs are processed by local councils and

regulated by jurisdictional (state and territory) governments. As a result, different jurisdictions have varying definitions and data processes for PAs. For example, the 'Planning Application (PA)' in Victoria (VIC) (Victorian Building Authority, 2023) is equivalent to the 'Development Application (DA)' and 'Construction Certificate' in New South Wales (NSW) (NSW Department of Planning and Environment, 2018). Moreover, the storage of these applications is disaggregated and in varying formats across local council websites and planning portals. Finally, PA descriptions are contained in unstructured written text formats which necessitates a methodology to process and compare this data across different sources.

Historically, these factors have limited the potential for data harmonisation between Australia PA data sources. Data harmonisation refers to the practice of combining different datasets to enhance their comparability or compatibility (Cheng et al., 2024). This may be achieved through a common identification key, shared features or a consistent classification scheme (Cheng et al., 2024). In the case of PAs, aggregating PA data into a single dataset has been limited by the complexity of extracting PA information from heterogeneous jurisdictional and council data sources. Moreover, disaggregated and inconsistent data collection practices across jurisdictions and local councils have made the use of identification keys or shared features untenable. Finally, while technically possible, applying a transferable classification scheme to compare datasets has historically been unfeasible due to the inefficiency and labour intensity of manually processing and analysing unstructured PA data.

Recently however, Natural Language Processing (NLP) models have emerged as a tool to aid urban planners with processing big textual data. These models enable the representation and analysis of human language computationally (Khurana et al., 2022). Advancements over the last decade in transformer based pre-trained models have made NLP tasks much more powerful and computationally efficient (Khurana et al., 2022). Urban analytics, loosely defined as the 'core set of tools employed to deal with problems of big data, urban simulation, and geodemographics' (Batty, 2018), has quickly adopted NLP techniques into common practice. In urban analytics, where planners regularly analyse unstructured text documents, NLP models have achieved significant progress in information and feature extraction, sentiment analysis, and classification tasks (Wu et al., 2022; Fu, 2024; Han et al., 2020).

This paper explores the potential for emerging NLP techniques to process and synthesise PA data. The paper implements a web scraping tool using Python – orchestrated in Amazon Web Services (AWS), consisting of 22 different individual web scrapers with targeted information extraction structures for different council websites (n=20) and government planning portals (n=2), covering 383 of 537 total councils, to harmonise planning text data from across Australia. The remaining councils were not included due to a lack of available data and resource limitations in developing web scraping tools. The

data is geocoded, filtered for ‘high priority’ applications, cleaned, and manually labelled into four different urban development categories to create a training dataset. The approach then assesses five different NLP models to predict the urban development category of a PA based on the description field. NLP models include supervised FastText and TextCNN, fine-tuned BERT, and unsupervised Word2Vec and ZSL models. Finally, the best-performing BERT model is applied to classify current PAs in Australia’s two largest cities, Sydney and Melbourne, to sense check the model classifications and evaluate current spatial trends of urban development in these cities.

The paper makes several key contributions to the field of urban analytics. Firstly, to our understanding, the paper presents the first approach to collect, geocode and analyse Australian PA data. While previous approaches have utilised structured datasets of city- or jurisdiction-wide PA data, our approach directly interacts with council websites and planning portals, enabling near real-time and comprehensive data collection at a national level. Secondly, the paper applies NLP techniques for the first time to develop a classification structure for the unstructured PA text descriptions. This creates a scalable and transferable approach to analyse broad urban development trends across previously restrictive spatial and temporal boundaries. Thirdly, the paper tests a range of supervised, fine-tuned, and unsupervised modelling approaches to provide insights into the suitability of different methods for planning text classification. The superior performance of the fine-tuned BERT models indicates that integrating pre-trained models with a supervised approach may be most accurate for PA classification. Finally, the paper makes empirical contributions by applying the NLP classification model to understand current housing pipeline trends in Sydney and Melbourne. Across both cities, the model reveals commercial development close to city centres, new housing and land developments towards specific spatial clusters in the urban fringes, and disaggregated clusters of alterations and modifications throughout the cities. These insights can help planners understand the responsiveness of the housing market to development-oriented policies and enable targeted timely infrastructure delivery for growing areas.

Holistically, the paper presents a novel and comprehensive approach to collect, process, and classify planning and urban development data to inform planning decisions.

2. Literature review

2.1 NLP in urban analytics

Urban planners regularly analyse documents containing large amounts of text data, such as land use plans and applications, reports, and policies (Fu, 2024). The process of digesting and analysing these documents into targetable insights is timely and subjective. Moreover, there is limited ability to

compare features of planning policies and documents across time and jurisdictions. NLP has emerged as a tool to aid planners with processing big textual data, significantly improving the efficiency and scalability of this process. While existing research is primarily exploratory, with limited practical applications, there have been several use cases emerge to leverage NLP in urban analytics (Fu, 2024).

Firstly, NLP is used in planning to extract key pieces of information such as dates, addresses, ID numbers, costs and other features from unstructured written text in websites, PDFs or other text documents (Wu et al., 2022). This approach has two advantages: first, it improves the efficiency of feature extraction, unlocking large amounts of previously underutilised data sources. (Tyagi and Bhushan, 2023). Second, it enables retrospective information extraction, overcoming hurdles of incorrect or inconsistent information structures during data collection. This approach has assisted in information querying from planning regulatory documents (Bommarito et al., 2021) and feature extraction from construction safety reports (Tixier et al., 2016).

Sentiment analysis is another prominent NLP application in the urban analytics field, performed primarily on media data (e.g., social media, print media) to evaluate public opinion towards major planning and development features (Fu, 2024). This NLP application has facilitated macro scale analysis of unstructured media sources, enabling a more direct evaluation of public opinion than traditional survey and interview methods. Use cases include assessing the perception of residents towards key neighbourhood characteristics through online reviews (Hu et al., 2019) and assessing public sentiment towards housing prices as evidenced through media articles (Biktimirov et al., 2024).

Finally, classification or semantic analysis is performed to classify unstructured text fields into categories or clusters (Fu, 2024). Classifying text documents has the advantage of enabling trend analysis and comparability of information which may have written formats or content. This capability can assist harmonisation of data sources with different feature definition or data collection structures (Cheng et al., 2024). In interpreting spatial plans, NLP has been applied to perform topic modelling on sustainability plans from the resilient cities network (Fu et al., 2022), and to identify planning topics in Canadian planning documents (Han et al., 2020). These studies implemented both unsupervised and supervised learning techniques to perform topic modelling on documents and analyse trends within the classified topics. These approaches have a high degree of transferability to the semantic analysis of PA urban development classes. One could fairly expect that the reasonably high accuracy and robustness achieved within these studies (e.g., 80% accuracy for Fu et al.'s (2022) topic modelling approach) could be achieved or improved upon within a PA context.

Yet, to our knowledge, there have been no attempts to apply NLP to analyse planning or building application documents. While a handful of studies have analysed structured PA datasets in relation to

housing supply and prices (e.g., Gabriel & Kung, 2024; Zhu et al., 2022), research analysing unstructured PA data is limited. While historically, the unstructured nature of PAs has limited detailed analysis, the emergence of NLP techniques for processing and analysing unstructured text data offers the potential for PA analysis to contribute timely and meaningful insights for urban planners and policymakers into the changing nature of our cities.

2.2 NLP model selection

There are a broad range of NLP methods that can be applied to text classification tasks. In the last ten years, there has been significant acceleration in the development of new and novel NLP methods. These methods may differ based on the vectorisation methods (method to convert raw text data into numerical format that machine learning algorithms can understand) used, degree of supervision, integration of Pre-trained Language Models (PLMs), and other factors. Different models will be suited to different classification tasks, dependent on the data, resources, and technical expertise available. This paper applies five models that span a range of complexity and data availability levels to help inform the optimal practical implementation approach of a PA classification model. Characteristics of the selected NLP models (Word2Vec, TextCNN, FastText, Bidirectional Encoders Representations from Transformers (BERT) and Zero-Shot Learning (ZSL)) are summarised in Table 1.

Table 1. Summary of selected NLP classification models, key model attributes and use cases in urban analytics.

Model	Year	Model type	Description	Urban analytics use cases
Word2Vec	2013	Unsupervised	Word2Vec is a word embedding technique that is used to transform individual words into a vector (Jatnika et al., 2019). Unsupervised Word2Vec models have been trained on online databases such as Wikipedia or Google News (Major et al., 2018).	• Classification of public complaints about city planning (Kim et al., 2021).
TextCNN	2014	Supervised	TextCNN employs a Convolutional Neural Network (CNN) adapted from traditional computer-vision models for text data, by applying a one-dimensional convolutional layer over pre-trained word embeddings such as Word2Vec (Kim, 2014).	• Classification of urban governance documents into governance categories (Wu et al., 2023).

FastText	2016	Supervised	FastText was developed as an extension of the Word2Vec architecture. Unlike Word2Vec, it considers word context and structure within sentences to learn more enriched word vectors (Bojanowski et al., 2017).	• Classification of geotagged urban environment labels (Hiippala et al., 2019). • Classification of social media review topics of transport planning (Sarram & Ivey, 2022).
BERT	2018	Supervised	BERT is a transformer-based PLM. It is a 24-layer Transformer with 340 million parameters that is pre-trained on English Wikipedia with 2.5 billion words (Devlin et al., 2019). For text classification tasks, the parameters of the pre-trained BERT model are frozen and only the output layer parameters are optimized (Sun et al., 2019). Modified BERT models such as DistillBERT and DeBERTA-V3 may offer additional performance improvements over the base BERT model (He et al., 2021).	• Mapping urban green space based on location name text descriptions (Cao et al., 2023). • Topic analysis on transport typology documents (Rath & Chow, 2022).
ZSL	2019	Unsupervised	Zero-Shot Learning (ZSL) is an unsupervised learning approach that transfers knowledge from previous training instances to classify testing sets, given seen classes (Yin et al., 2019). ZSL can be treated as a natural language inference problem, leveraging PLMs such as BERT as an entailment-based ZSL text classifier (Devlin et al., 2019).	• Topic analysis on smart city documents (Lim & Hwang, 2024).

3. Methods and Data

3.1 Australian PA process

In Australia, PAs are collected by various jurisdictional and council authorities for consultation with the community. Although each local council has its own planning scheme and requirements for the application process, they are generally guided by overarching jurisdictional government planning

schemes that regulate the use and development of land. The regulations and naming conventions for PAs vary across jurisdictions, outlined in Table 2. All open PAs must be publicly listed to allow community consultation, which may occur through letters, print media, or council websites. In most cases, data is recorded on a public, albeit generally unstructured, online source. This provides an opportunity, with appropriate tools and capabilities, to access and process this data for analysis.

Table 2. Information of planning applications data harmonised across Australia jurisdictions.

Jurisdiction	PA title	Decision-maker	Decision time (business days)
New South Wales (NSW)	Development Application	LGA	**40 - 90 days** (Section 91 of New South Wales Environmental Planning and Assessment Regulation 2021).
	Complying Development Certificate	LGA or accredited certifier	**10 - 20 days** (Section 133 of New South Wales Environmental Planning and Assessment Regulation 2021).
Victoria	Planning Permit	LGA	**28 - 49 days** (Section 47 of Victoria Planning and Environment Act 1987).
Queensland	Development Application	LGA	**35 days** (Development Assessment Rules, Queensland Department of State Development, Infrastructure, Local Government and Planning 2021).
South Australia	Development Application	LGA or relevant authorities	**5 - 95 days** (Section 53 of South Australia Planning, Development and Infrastructure Act 2016).
Western Australia	Development Application	LGA	**60 - 90 days** (Section 75 of West Australia Planning and Development (Local Planning Schemes) Regulations, 2015).
Tasmania	Planning Permit	LGA	**14 - 21 days** (Section 54 of Tasmania Land Use Planning and Approvals Act 1993).
Northern Territory	Development Permit	Development consent authority	**14 – 30 days** (Part 5 of Northern Territory Planning Act 1999).

3.2 PA data processing

The process to harmonise PA data from different jurisdictions consisted of an automated web scraping, geocoding and data filtering pipeline. The pipeline was tailored to function on 22 different online data sources (listed in Appendix A), navigating differences in data formats and PA definitions across local council websites and government planning portals. The pipeline consisted of Python code, executed

using cloud computing services through AWS, run daily to provide a near real-time dataset of open PAs in Australia.

Firstly, a web scraper tool was developed in Python to scrape PA information stored in schemes such as JSON or table-like formats. The tool scraped information from local council websites, the planning portal PlanningAlert, and the NSW Government website. The scraper was built using the Python packages BeautifulSoup and Selenium to extract key pieces of information such as address, application date, permit type, and application description. The scraper identified these pieces of information by scanning for key headings or labels in the online data sources. Secondly, the data was geocoded by applying an AWS geocoder on the scraped address field. Finally, keywords were identified in the PA description field to classify the observations into 'low' and 'high priority' applications, based on the impact of each PA on urban activity within the development area. 'High priority' keywords included ["dwelling", "office", "redevelopment", "subdivision"] while low priority keywords included ["paint", "fence", "door", "repair"]. PAs that consisted of at least one 'high priority' keyword and no 'low priority' keywords were labelled as 'high priority' PAs.

The output of the automated pipeline, executed using AWS, was a daily dataset containing all current, 'high priority' PAs scraped from available council and planning online data sources. As of 15 August 2024, this dataset contained 378,278 planning records collected from 383 councils (out of 537 total councils in Australia).

3.3 PA classification methodology

The output PA dataset from the pipeline was then cleaned, labelled, and used to train several NLP classification models. The model was used to classify PAs into four distinct urban development categories based on the PA description field: "Commercial development", "New housing development", "Subdivision and land development", and "Alterations and modifications". The best performing model was applied to map urban development trends in Sydney and Melbourne.

As a first step for the classification model, the text in the description field for each PA was cleaned and converted to numerical representation for use with the NLP models. The steps undertaken to prepare the description field are outlined in the supplementary materials.

The cleaned series of vectorised descriptions comprised the training dataset. Manual labelling was required to generate labels for the supervised learning models. A random sub-sample of 3,000 observations from Sydney were manually labelled into the four urban development categories. Sydney was chosen as the training dataset area, given the known diversity and density of current urban development, while testing indicated that 3,000 training data points was sufficient to generate

robust modelling results. The model was split into a training / test / validation dataset using a 60 / 20 / 20 data split.

The model was then built using the cleaned and vectorised description fields to predict the urban development category labels. Five different model architectures were tested: Word2Vec, TextCNN, FastText, BERT, and ZSL using a BERT base layer. For models that contained a pre-trained component, several different PLM variants were tested, as reported in section 4.1.

Finally, the best performing model (DeBERTa-v3) was applied to classify current PAs in Sydney and Melbourne into one of the four urban development categories. The model was applied to PA data from two time periods in both cities to assess changes in the spatial trends of urban development. The data was first scraped for PA's available as of 1st January 2023, containing open PAs lodged from 2021-2022 (2021-2022). The second dataset was scraped as of 20th August 2024 and filtered to contain PAs lodged from 2023 onwards (2023-2024). In NSW, the first and second datasets contained 66,057 and 84,519 observations respectively, while in Victoria the two datasets contained 16,471 and 26,641 observations respectively. NSW contains a greater amount of PA data due to the NSW government planning portal centralising PA data sources, whereas Victoria lacks such a portal. The spatial distribution of urban development clusters in both cities was mapped using Kepler.gl, a GPU accelerated front-end tool which provides high performance visual analytics (Soundararaj & Pettit 2024). While the model was not applied to other Australian cities or regional areas, the PA data is downloaded, and future work could run the classification nationally with minimal additional effort.

4. Results and Discussion

4.1. Model performance

The experimental results of different methods for planning classification are summarized in Table 3. The accuracy, which measures the proportion of correct predictions, and the F1 score, which measures the balance between precision and recall, are reported. These measures provide an indication of the absolute and relative accuracy of the model across the four classification classes. Because the classification task aims to assess the housing pipeline across all four development categories, with no heightened cost for correctly or incorrectly classifying a particular class, these general performance metrics were deemed sufficient to assess the suitability of competing models.

The first and most notable distinction across the models is the significant difference in performance between supervised and unsupervised models. The best-performing unsupervised model, Word2Vec Google News, significantly underperforms the worst-performing supervised model, FastText, with an

accuracy of 0.760 compared to 0.878. Despite considerable testing of variations in class labels and PLM architectures, we were unable to achieve competitive accuracy for the unsupervised learning models. Potential causes for this underperformance may include limited information to form classifications, with many PA descriptions containing less than 10 words, and insufficient semantic variation in the class labels, with frequent misclassifications occurring between 'new housing construction' and 'alterations and modifications'.

Table 3. Accuracy and F1-scores of tested text classification models.

Model type	Model	PLM (if applicable)	Accuracy	F1-Score
Supervised	**TextCNN**		0.904	0.900
	FastText		0.878	0.877
	BERT	Bert-base-uncased	0.912	0.912
		Distilbert-base-uncased	0.932	0.931
		Nli-deberta-v3-base	**0.944**	**0.944**
Unsupervised	**Word2Vec**	Glove-wiki-gigaword-300	0.675	0.651
		Word2Vec-google-news-300	0.760	0.750
	ZSL	Bert-base-uncased	0.631	0.650
		Nli-deberta-v3-base	0.649	0.671

Among supervised models, the performance of the fine-tuned BERT models is better than the self-taught FastText and TextCNN models, with accuracies ranging from 0.912-0.944 compared to 0.878 and 0.904, respectively. This indicates the effectiveness of leveraging knowledge from other English text datasets for planning text classification. DeBERTA-v3 method is the best-performing BERT model across both metrics, with an accuracy and F1-score of 0.944. This aligns with literature documenting the performance improvements of modified BERT models such DeBERTA-v3 over the base BERT model (He et al., 2021).

Overall, the model performance across a diverse range of NLP models indicates that for a bespoke planning text classification task, with semantically similar classes and potentially limited text data for some observations, fine-tuned models may be best suited to leverage PLM components while considering the specific planning text context.

4.2 Urban development clusters

The best performing DeBERTA-v3 model was applied to map PA classifications over the two time periods (2021-2022, 2023-2024) in Sydney and Melbourne. Understanding spatial and temporal urban

development trends can perform a basic sense check of the classification model and provide insights on the housing supply pipeline across both cities.

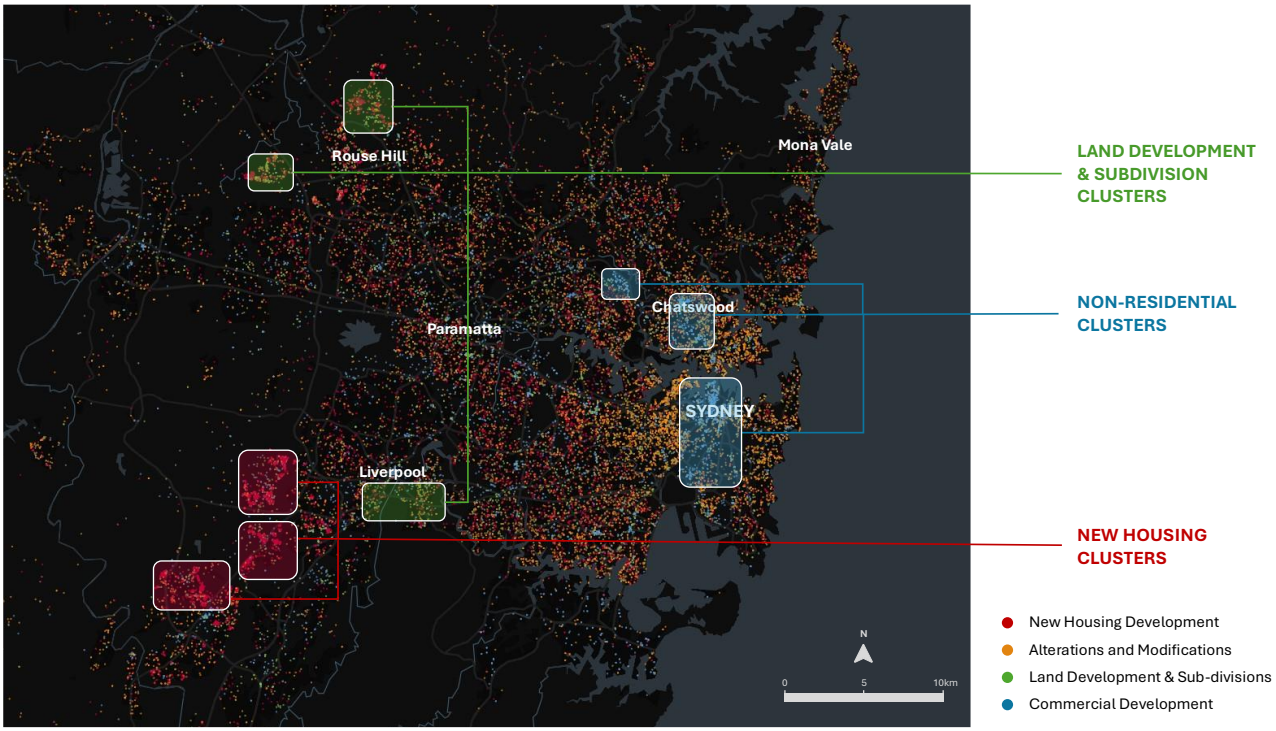
An analysis first of Sydney's 2021-2022 map (Figure 1 (a)) reveals several spatial clusters of urban development activity which support a high-level face-validation of the classification model. This includes new housing construction to the south-west of the city, in new land release areas towards the Badgerys Creek Airport expansion area (Koziol, 2020). In greenfield sites there are spatial clusters of new land development subdivisions to the north-west and south-west, while there are several non-residential clusters towards the traditional business districts of the Sydney CBD and Chatswood City. Finally, there are alterations and modifications to the existing urban fabric occurring throughout the city as it experiences urban renewal.

Investigating PAs from 2023-2024 reveals several new urban development clusters (Figure 1 (b)). Firstly, there is a major commercial development cluster around Parramatta City, reflecting its emergence as Sydney's 'second CBD' (Bolger & Bowring, 2024). South of Parramatta City, new housing construction is clustered around the suburbs of Auburn and Bankstown, indicating property growth and potential gentrification in this area (Thackway et al., 2023b). There is an additional subdivision cluster towards greenfield sites in the south-west, while previous subdivision, new housing, and commercial development clusters remain in the north-west, south-west, and city centre respectively. Interestingly, the map reveals a drop-off in alterations and modifications throughout the city, particularly in the inner west. This may reflect increases to construction costs due to global supply chain issues, and/or the erosion of households' borrowing capacity due to high interest rates (Johanson, 2024).

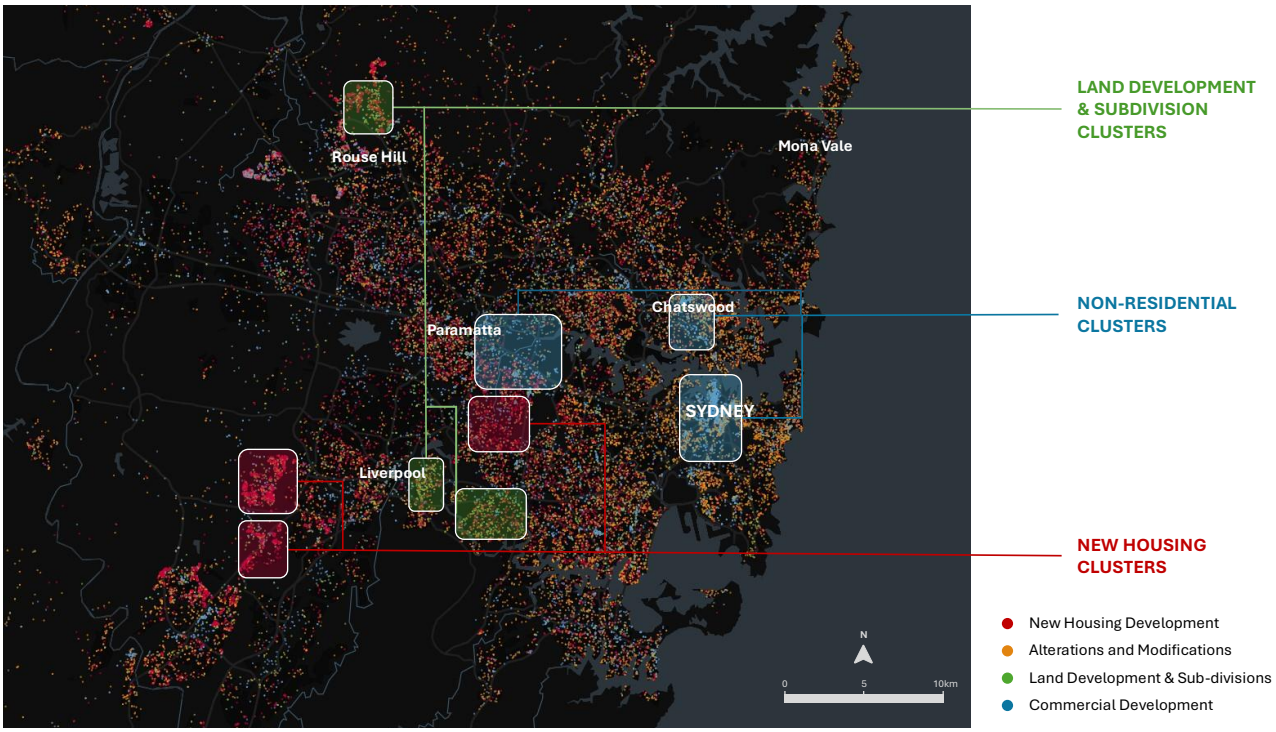
Sydney's development profile raises several interesting considerations for planners. Firstly, do PAs between the two time periods indicate an uptick in the housing supply pipeline, which could potentially assist in improving housing affordability (NSW Government Planning, 2024)? In NSW, new house construction PAs classified by our model increased between January 2023 (19,012) and August 2024 (25,225), however this figure is still well below the annual target of 45,200 new homes set by NSW Government Planning (NSW Government Planning, 2024). Tracking new housing PAs in the coming years will be particularly noteworthy to evaluate NSW and Australia's responsiveness to Australia's 'National Housing Accord' which aims to build 1.2 million new homes by mid-2029 (Australian Government Treasury, 2024). Secondly, are growth areas sufficiently supported by existing or new transport and service infrastructure? In Sydney, the corridors of commercial and land development towards Parramatta, Rouse Hill, and the Badgerys Creek Airport are directly serviced by Sydney's new Metro train line (Sydney Metro, 2024), indicating alignment between planning and new transformational transport infrastructure, both metro and airplane.

Figure 1. PA urban development classifications by NLP model (DeBERTA-v3) in Sydney.

a) 2021-2022, as of January 2023.



b) 2023-2024, as of August 2024.



In Melbourne, while there is less PA data, similar trends are evident in the spatial distribution of urban development. Between 2021-2022 (Figure 2 (a)), new housing construction clusters reflect residential growth in Melbourne's 'booming west' (Development Victoria, 2024). There are several land development clusters in outer East Melbourne where planning regulations are generally more

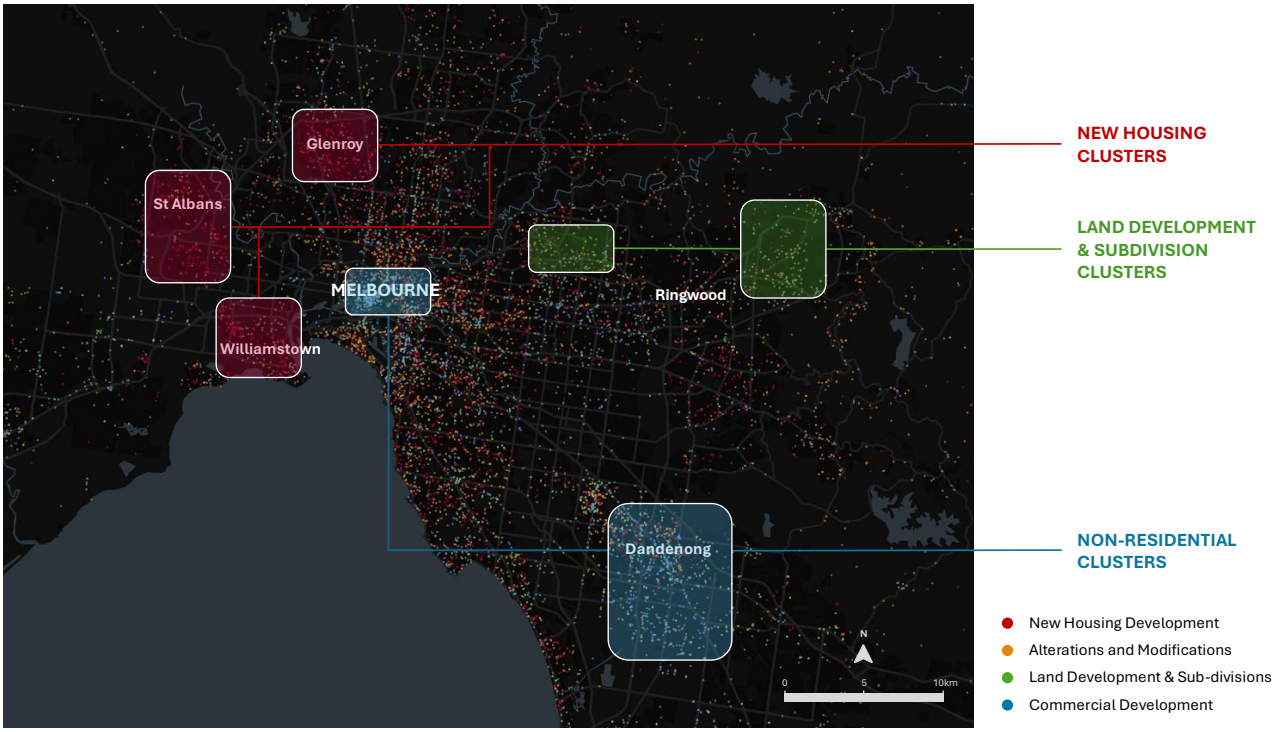
favourable to subdivisions (Robbins, 2022). Finally, there is dense commercial development in Melbourne's CBD and towards Dandenong, indicating the success of a 2021 Victorian Government initiative to stimulate commercial growth in the Dandenong area by reducing development taxes (Development Victoria, 2021).

Between 2023-24, while the prevailing development patterns remain similar, there are several emerging clusters (Figure 2 (b)). The Victorian Government development incentive in Dandenong has stimulated both non-residential and subdivision activity in the area, while there is a new subdivision cluster towards St Albans where development was previously more focused on new housing construction. Additionally, there is significant commercial activity south-east of the Melbourne CBD which likely reflects planning for the 'South Yarra Square' commercial precinct project which emerged in late 2022 (Schlesinger, 2022).

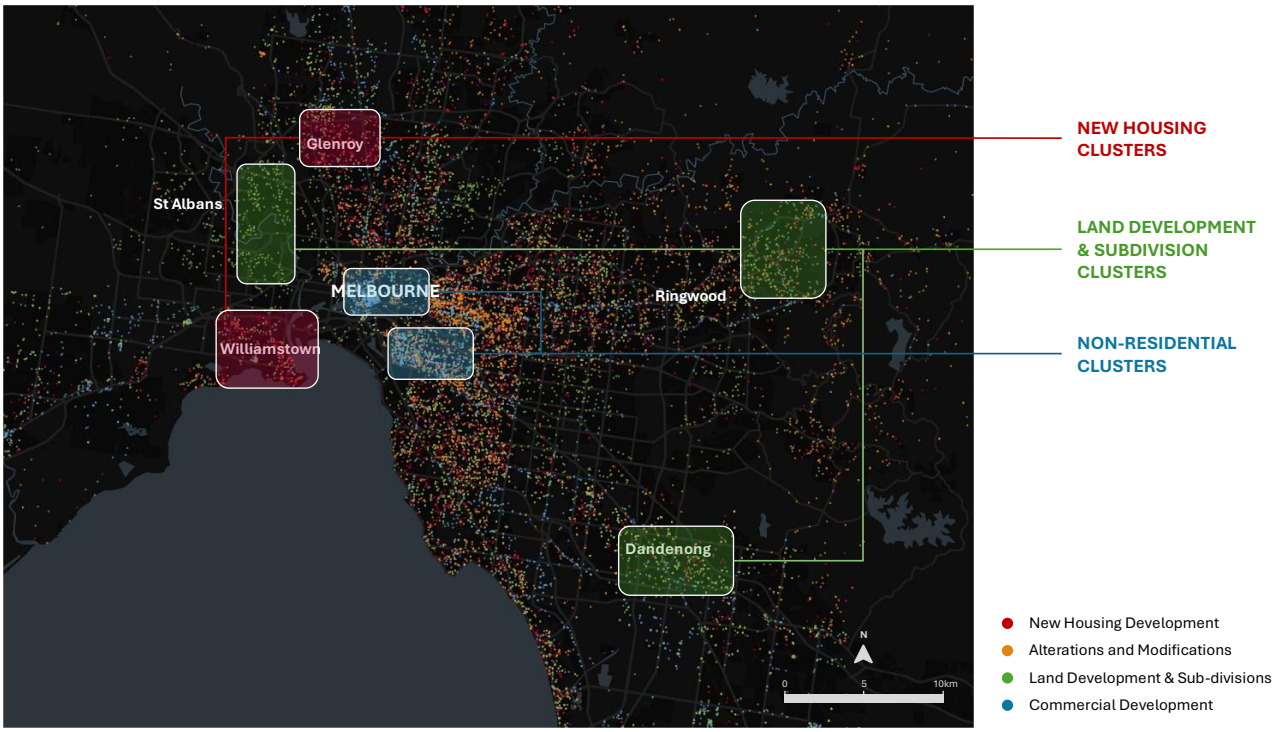
In Victoria, as in NSW, new housing construction PAs have gone up slightly (4,494 in January 2023, 5,564 in August 2024), however these figures may be less complete given the sparser web scraped data for Victoria. Planners may also be interested in the positive market responses to government initiatives to stimulate commercial and land development in South Yarra and Dandenong, although this growth must be supported by sufficient services and transport infrastructure.\

Figure 2. PA urban development classifications by NLP model (DeBERTA-v3) in Melbourne.

a) 2021-2022, as of January 2023.



b) 2023-2024, as of August 2024.



Across both metropolitan areas, similar spatial trends are evident, with new housing and subdivisions in the outer suburbs, non-residential development towards the city centres, and alteration and modifications distributed throughout. These overall trends provide a face-validation of the NLP model's classifications. Moreover, investigating PA classifications across time and space illuminates several urban development trends, such as emerging commercial districts and new housing hotspots. Such insights can support planners to better understand the development industry's responsiveness to policy initiatives and the planning needs for future growth areas.

5. Conclusions

PAs have recently garnered attention as a rich and publicly available data source to understand the future housing supply pipeline and infrastructure requirements of an area (Zhu et al., 2022; Murray, 2023). PAs under consideration are often required to be listed for public information and consultation and are generally available through council websites and government planning portals. However, previous studies have been unable to directly interact with PAs due to difficulties reconciling differences in jurisdictional PA definitions, online data formats, and an inability to process large amounts of unstructured text data. Our study implements large scale web-scraping, geocoding, data harmonisation, and emerging NLP capabilities to develop a novel pipeline that processes, classifies and maps PAs to create a robust and scalable method to analyse PAs on a metropolitan scale.

Firstly, our approach implements a customised web-scraping tool, built in Python and executed daily using AWS, to scrape current PAs from 22 different online data sources. The process applied a

geocoder and a ‘priority’ classification using keywords to provide a daily, geocoded dataset of ‘high priority’ PAs under consideration nationally. The PA dataset was labelled and classified into four classes, differentiated by the infrastructure provisioning requirements of each development category. To our knowledge, this is the first approach to collect and harmonise PA into a nationally consistent classification schema.

The classification approach tested five different NLP models that integrated varying degrees of pre-trained models in supervised and unsupervised contexts. While the unsupervised PLMs performed poorly, the best performing method was the fine-tuned DeBERTA-v3 model, with an accuracy and F1-score of 0.944. This indicates that fine-tuning a supervised PLM can maximise performance for planning text classification tasks. These methodological findings can help inform future NLP modelling approaches within the urban analytics field.

Finally, the fine-tuned DeBERTA-v3 model was applied to classify and map urban development trends in Sydney and Melbourne over two time periods: 2021-2022 and 2023-2024. The mapping exercise revealed broadly consistent development trends across both cities and time-periods, providing a face-validation of the classification model. Moreover, the classifications provided insights into the new housing supply pipeline, which displayed a slight increase in both cities, and the different types of urban growth areas. While the appropriateness of infrastructure provisioning in growth areas could be evaluated to some degree through a high-level assessment of new transport connections, further analysis of economic and service provisions in these areas could support a deeper understanding. Nonetheless, the classification model provides a highly accurate and robust method to understand new housing supply and urban growth area typologies.

Future research could look to expand the comprehensiveness of the PA dataset by incorporating additional council websites and aggregated planning datasets into the web-scraping approach. The classification model could also be extended, potentially with additional PA information such as expected cost and council location, to perform predictive modelling such as expected PA outcomes or timelines. Further work could also include assessments of infrastructure delivery in commercial and new build areas to assess the effectiveness of planning in growth areas.

Notwithstanding this, the study provides the first approach to harmonise and analyse PA data on a metropolitan scale. This approach utilises data and modelling advancements to aggregate a national PA dataset and deliver a robust urban development classification schema. Ultimately, the study demonstrates the potential for machine learning methods to automate PA analysis and provide urban analytic driven insights that better inform strategic planning and policymaking.

References

Australian Government Treasury, 2024. *Delivering the National Housing Accord*. Available at: https://treasury.gov.au/policy-topics/housing/accord (accessed 26 August 2024).

Batty, M., 2019. Urban analytics defined. *Environment and Planning B: Urban Analytics and City Science*, 46(3), doi: https://doi.org/10.1177/2399808319839494.

Bolger, R., and Bowring, D., 2024. *Parramatta rapidly transformed into Sydney's second CBD. Hopes are high for the next phase*. May 2024, ABC News, available at: https://www.abc.net.au/news/2024-05-01/parramatta-changes-dramatically-in-ten-years/103779398 (accessed 23 August 2024).

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T., 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146, doi: https://doi.org/10.1162/tacl_a_00051.

Bommarito, I.I., Katz, M.J., Detterman, E.M., 2021, LexNLP: natural language processing and information extraction for legal and regulatory texts. *Research Handbook on Big Data Law*, Edward Elgar Publishing.

Biktimirov, EN., Sokolyk, T., and Ayanso, A., 2024. What is behind housing sentiment? *Finance Research Letters*, 60, 104966, doi: https://doi.org/10.1016/j.frl.2023.104966.

Caliyurt, O., 2022. The Mental Health Consequences of the Global Housing Crisis. *Alpha Psychiatry*, 23(6), pp. 264-265, doi: 10.5152/alphapsychiatry.2022.17112022.

Cao, S., Zhao, X., and Du, S., 2023. Multi-type and fine-grained urban green space function mapping based on BERT model and multi-source data fusion. *International Journal of Digital Health*, 17(1), 2308723, doi: https://doi.org/10.1080/17538947.2024.2308723.

Cheng, C., Messerschmidt, L., Bravo, I., Waldbauer, M., Bhavikatti, R., Schenk, C., Grujic, V., Model, T., Kubinec, R., and Barcelo, J., 2024. A General Primer for Data Harmonization. *Scientific Data*, 11, 152, doi: https://doi.org/10.1038/s41597-024-02956-3.

Development Victoria, 2021. *Boost for investment in Central Dandenong*. April 2021, Development Victoria, available at: https://www.development.vic.gov.au/news/boost-for-investment-in-central-dandenong (accessed 23 August 2024).

Development Victoria, 2024. *More new homes on the way for Melbourne's booming west*. April

2024, Development Victoria, available at: https://www.development.vic.gov.au/news/more-new-homes-on-the-way-for-melbournes-booming-west (accessed 23 August, 2024).

Devlin, J., Chang, M-W., Lee, K., and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv preprint*, doi: https://doi.org/10.48550/arXiv.1810.04805.

Fauth, J., and Soibelman, L., 2022. Conceptual Framework for Building Permit Process Modeling: Lessons Learned from a Comparison between Germany and the United States regarding the As-Is Building Permit Processes. *Buildings*, 12(5), 638, doi: https://doi.org/10.3390/buildings12050638.

Ferm, J., Clifford, B., Canelas, P., and Livingstone, N., 2021. Emerging problematics of deregulating the urban: The case of permitted development in England. *Urban Studies*, 58(10), pp. 2040-2058, doi: https://doi.org/10.1177/00420980209369.

Fu X. (2024), Natural Language Processing in Urban Planning: A Research Agenda. *Journal of Planning Literature*, 0(0), doi: https://doi.org/10.1177/08854122241229571.

Fu, X., Li, C., and Zhai, W., 2022. Using Natural Language Processing to Read Plans: A Study of 78 Resilience Plans From the 100 Resilient Cities Network. *Journal of the American Planning Association*, *89*(1), pp. 107–119, doi: https://doi.org/10.1080/01944363.2022.2038659.

Gabriel, S., and Kung, E., 2024. *Development Approval Timelines, Approval Uncertainty, and New Housing Supply: Evidence from Los Angeles*. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4872147.

Glaeser, E., and Gyourko, J., 2018. The Economic Implications of Housing Supply. *Journal of Economic Perspectives*, 32(1), pp. 3-30, doi: 10.1257/jep.32.1.3.

Gunn, L., Kroen, A., Gruyter, C.D., Higgs, C., Saghapour, T., and Davern, M., 2020. Early delivery of equitable and healthy transport options in new suburbs: Policy, place and people. *Journal of Transport & Health*. 18, 100870, doi: https://doi.org/10.1016/j.jth.2020.100870.

Han, A.T., Laurian, L., and Dewald, J., 2020. Plans Versus Political Priorities: Lessons From Municipal Election Candidates' Social Media Communications. *Journal of the American Planning Association*, *87*(2), pp. 211–227, doi: https://doi.org/10.1080/01944363.2020.1831401.

He, P., Gao, J., and Chen, W., 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *ArXiV preprint*, doi: https://doi.org/10.48550/arXiv.2111.09543.

Hiippala, T., Hausmann, A., Tenkanen, H., and Toivonen, T., 2019. Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*, 34(2), pp. 290-309, doi: https://doi.org/10.1093/llc/fqy049.

Hu, Y., Deng, C., and Zhou, Z., 2019. A Semantic and Sentiment Analysis on Online Neighborhood Reviews for Understanding the Perceptions of People Toward Their Living Environments. *Annals of the American Association of Geographers*, 109(4), pp. 1052–73, doi: https://doi.org/10.1080/24694452.2018.1535886.

Johanson, S., 2024. *High interest rates erode households' ability to borrow and build*. Sydney Morning Herald, March 2024, available at: https://www.smh.com.au/business/companies/high-interest-rates-erode-households-ability-to-borrow-and-build-20240315-p5fcom.html (accessed 17 September 2024).

Kellet, J., and Nunnington, N., 2019. Infrastructure for new Australian housing: Who pays and how? *Cities*, 92, pp. 10-17, doi: https://doi.org/10.1016/j.cities.2019.03.007.

Khurana, D., Koli, A., Khatter, K., and Singh, S., 2022. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools And Applications*, 82, pp. 3713-3744, doi: https://doi.org/10.1007/s11042-022-13428-4.

Kim, Y., 2014. Convolutional neural networks for sentence classification. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October. 1746–1751.

Koziol, M., 2020. *New York, London, Paris, Badgerys Creek: The developers' gold mine on Sydney's fringe*. Oct 2020, Sydney Morning Herald, available at: https://www.smh.com.au/national/nsw/new-york-london-paris-badgerys-creek-the-developers-gold-mine-on-sydney-s-fringe-20201021-p567cz.html (accessed 23 August 2024).

Lim, J., and Hwang, J., 2024. Exploring diverse interests of collaborators in smart cities: A topic analysis using LDA and BERT. *Heliyon*, 10(9), e30367, doi: https://doi.org/10.1016/j.heliyon.2024.e30367.

Major, V., Surkis, A., and Aphinyanaphongs, Y., 2018. Utility of General and Specific Word Embeddings for Classifying Translational Stages of Research. *AMIA Annu. Symp. Proc.*, 2018, pp. 1405-1414, PMID: 30815185.

Murray, C.K., 2023. *Explainer: Planning and building approvals*. Henry Halloran Research Trust, The University of Sydney, doi: 10.31219/osf.io/8tb7v.

New South Wales Legislation, 2024. *Environmental planning and assessment regulation 2021*. Available at: https://legislation.nsw.gov.au/view/html/inforce/current/sl-2021-0759 (accessed 15 Aug 2024).

Northern Territory Legislation, 2024. *Planning Act 1999*. Available at: https://legislation.nt.gov.au/Legislation/PLANNING-ACT-1999 (accessed 18 September 2024).

NSW Department of Planning and Environment, 2018. *Your guide to the development application process*. Available at: https://www.planningportal.nsw.gov.au/sites/default/files/documents/2019/da- bestpractice-guide-for-homeowners-2018-06-07.pdf (accessed 01 June 2023).

NSW Government Planning, 2024. *Demand for housing in NSW*. Available at: https://www.planning.nsw.gov.au/policy-and-legislation/housing/housing-supply-insights/quarterly-insights-monitor-q1/demand-for-housing-in-nsw#:~:text=We%20will%20need%20an%20additional,pre%2DCOVID%20levels%20by%202025 (accessed 26 August 2024).

Phibbs, P., and Gurran, N., 2021. The role and significance of planning in the determination of house prices in Australia: Recent policy debates. *Environment and Planning A: Economy and Space*, *53*(3), pp. 457-479, doi: https://doi.org/10.1177/0308518X21988942.

Queensland Department of State Development, Infrastructure, Local Government and Planning, 2024. *Development assessment rules guidance for development assessment*. Available at: https://dsdmipprd.blob.core.windows.net/general/da-rules-guidance.pdf. (accessed 15 Aug 2024).

Rath, S., and Chow, J.Y.J., 2022. Worldwide city transport typology prediction with sentence-BERT based supervised learning via Wikipedia. *ArXiv prepint*, doi: https://doi.org/10.48550/arXiv.2204.05193.

Robbins, K., 2022. *Why subdivisions should be on the agenda if you own property in these 10 suburbs*. October 2022, Smart Property Investment, available at: https://www.smartpropertyinvestment.com.au/hotspots/24168-why-subdivisions-should-be-on-the-agenda-if-you-own-property-in-these-10-suburbs (accessed 23 August 2024).

Sarram, G., and Ivey, S.S., 2022. Evaluating the potential of online review data for augmenting traditional transportation planning performance management. *Journal of Urban Management*, 11(1), pp. 123-136, doi: https://doi.org/10.1016/j.jum.2022.01.001.

Schlesinger, L., 2022. *Plans lodged for $250m hotel and office project on Toorak Road*. October 2022, Australian Financial Review, available at: https://www.afr.com/property/commercial/plans-lodged-for-250m-hotel-and-office-project-on-toorak-road-20221004-p5bmy3 (accessed October 4 2022).

Shahzad, W.M., Hassan, A., and Rotimi, J.O.B., 2022. The challenges of land development for housing provision in New Zealand. *J Hous Built Environ*, 37(3), pp. 1319-1337, doi: 10.1007/s10901-021-09896-z.

Soundararaj, B., and Pettit, C., 2024. Australian property market explorer – a front-end native tool for visualizing property sales data. *Architectural Science Review*, doi: https://doi.org/10.1080/00038628.2024.2369658.

South Australia Legislation, 2024. *Planning, development and infrastructure act 2016*. Available at: https://www.legislation.sa.gov.au/lz?path=%2FC%2FA%2FPLANNING%20DEVELOPMENT%20AND%20INFRASTRUCTURE%20ACT%202016 (accessed 15 Aug 2024).

Sun, C., Qiu, X, Xu, Y., and Huang, X., 2019. How to fine-tune bert for text classification? In: *Chinese Computational Linguistics*, pp. 194–206, doi: https://doi.org/10.1007/978-3-030-32381-3_16.

Sydney Metro, 2024. *Sydney Metro Interactive Train Map*. Available at: https://www.sydneymetro.info/map/sydney-metro-interactive-train-map (accessed 26 August 2024).

Tasmanian Legislation, 2024. *Land use planning and approvals act 1993*. Available at: https://www.legislation.tas.gov.au/view/html/inforce/current/act-1993-070 (accessed 15 Aug 2024).

Thackway, W., Ng, M., Lee, C-L., and Pettit, C., 2023a. Implementing a deep-learning model using

Google street view to combine social and physical indicators of gentrification. *Computers, Environment and Urban Systems*, 102, 101970, doi: https://doi.org/10.1016/j.compenvurbsys.2023.101970.

Thackway, W., Ng, M., Lee, C-L., and Pettit, C., 2023b. Building a predictive machine learning model of gentrification in Sydney. *Cities*, 134, 104192, doi: https://doi.org/10.1016/j.cities.2023.104192.

Tixier, A.J-P., Hallowell, M.R., Rajogopalan, B., and Bowman, D., 2016. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62, pp. 45-56, doi: https://doi.org/10.1016/j.autcon.2015.11.001.

Tyagi, N., and Bhushan, B., 2023. Demystifying the Role of Natural Language Processing (NLP) in Smart City Applications: Background, Motivation, Recent Advances, and Future Research Directions. *Wirel Pers Commun*, 130(2), pp. 857-908, doi: 10.1007/s11277-023-10312-8.

Uddin, F.U., Piracha, A., and Phibbs, P., 2022. A tale of two cities: Contemporary urban planning policy and practice in Greater Sydney, NSW, Australia. *Cities*, 123, 103583, doi: https://www.sciencedirect.com/science/article/pii/S0264275122000221.

Van der Kooi, M., 2024. *The influence of issued building permits on housing price.* B. Thesis, University of Groningen, available at: https://frw.studenttheses.ub.rug.nl/id/eprint/4508 (accessed 17 September 2024).

Victorian Building Authority, 2023. *Planning and building permits*. Available from: https://www.vba.vic.gov.au/consumers/home-renovation-essentials/permits (accessed 15 Aug 2024).

Victorian Legislation, 2024. *Planning and environment act 1987*. Available at: https://www.legislation.vic.gov.au/in-force/acts/planning-and-environment-act-1987/156 (accessed 15 Aug 2024).

Western Australia Legislation, 2024. *Planning and development (local planning schemes) regulations 2015*. Available at: https://www.wa.gov.au/government/document-collections/planning-and-development-local-planning-schemes-regulations-2015 (accessed 15 Aug 2024).

Wang, F., and Niu, F-q., 2019. Urban Commercial Spatial Structure Optimization in the Metropolitan Area of Beijing: A Microscopic Perspective. *Sustainability*, 11(4), 1103, doi: https://doi.org/10.3390/su11041103.

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., and Yang, Z., 2022. Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, 134, 104059, doi: https://doi.org/10.1016/j.autcon.2021.104059.

Wu, J., Li, T., and Zheng, X., 2023. Urban Governance Information Classification Method Based on BERT-TextCNN. In *Proceedings of the 2nd International Conference on Mathematical Statistics and Economic Analysis*, Nanjing, China, doi: 10.4108/eai.26-5-2023.2334281.

Yin, W., Hay, J., and Roth, D., 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *ArXiv preprint*, doi: https://doi.org/10.48550/arXiv.1909.00161.

Zhu, Z., Shirowzhan, S., and Pettit, C.J., 2022. Investigation of development applications: A GIS based spatiotemporal analysis in the city of Sydney area 2004–2022. *Buildings*, 12 (10), 1601, doi: https://doi.org/10.3390/buildings12101601.

Appendix A. List of web-scrapers

Jurisdiction	**Data Source**
Australia	Planning Alerts
NSW	NSW Planning Portal
Queensland	Queenscliffe council
South Australia	Glenelg council
Tasmania	Burnie council
	Devonport council
	Glenorchy council
Victoria	Armadale council
	Colac Otway Shire council
	Dandenong council
	Indigo council
	Melton council
	Mornington council
	Warrnambool council
	Wodonga council

WA	Bayswater council
	Cambridge council
	Cottesloe council
	Fremantle council
	Kalamunda council
	Nedlands council
	Rockingham council