**Comparative Analysis of Cultural Disparities in Entertainment Preferences and Online Discourse on Local Food Culture between Melbourne and Sydney**

**Submitted to,**

**Jeffrey Chan**

**Associate Professor**

**Artificial Intelligence**

**RMIT University**

**Submitted by,**

**Group Name – Peak Performers**

**Members**

**Md Arafat Hossain – s4032285**

**Mashuk Arefin Pranjol – s4027801**

**Mohammed Manzoor - s3914638**

**Contents**

**Introduction 3**

**Data Collection 3**

**Pre-processing and Data Cleaning 3**

**Initial Exploration 4**

**Analysis approach 5**

**Network Graph 6**

**Topic Modelling 7**

**Sentiment Analysis 10**

**Conclusion 11**

**References 12**

**Introduction**

In an increasingly interconnected world, understanding the dynamics of cultural nuances and preferences is critical for effective communication. This study explores into the contrasts in entertainment inclinations and the online discussions surrounding local food cultures in two of Australia's most vibrant cities, Melbourne, and Sydney. Through a comprehensive social media analysis, this report aims to explore the underlying trends, sentiments, and topic distributions on the distinctive cultural landscapes within these urban communities.

**Data Collection**

For the data collection on Reddit, we have specifically chosen certain subreddits for Melbourne and Sydney. Once the subreddits were selected, we executed a daily code from 1 October to 19 October using the PRAW API. Reddit's API enables us to gather a substantial amount of data, hence this approach was necessary. Subsequently, all the collected data was consolidated into separate CSV files for Melbourne and Sydney. Overall, we amassed a total of 18,790 Reddit posts along with their corresponding comments.

Additionally, for an alternative data source, we opted for a Twitter dataset from Kaggle named "Australian Cities – Tweets". This dataset encompasses 138,495 tweets originating from various cities across Australia. This dataset is created in the July of 2020. From this dataset, we exclusively extracted tweets originating from Melbourne and Sydney each 10,000 tweet, which amounted to 20,000 in total.

**Pre-processing & data cleaning**

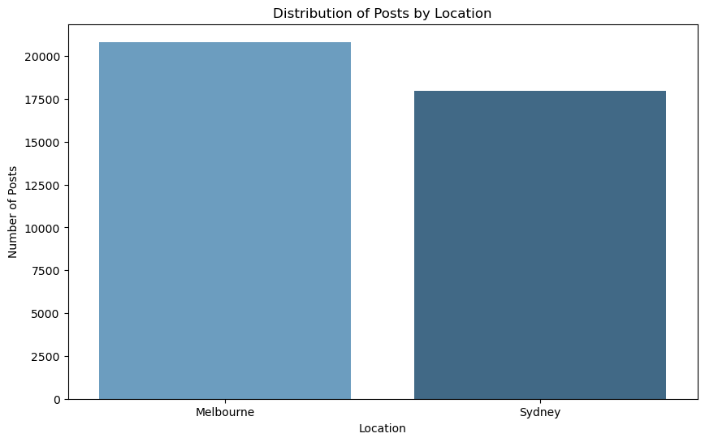
In our tweeter dataset there was a lot of redundant data columns. At first, we analysed our data by printing the head of the dataframe. After that, we dropped all these redundant columns. We also had to match the time format with our reddit dataset. We also have changed the heading name to match our reddit dataset. After that we have merged all the tweeter and reddit post from Melbourne and Sydney in a single csv file.

In our approach, we removed any non-English text from the title and comment column because we will be analysing this on English language. We also removed any URLs, user mentions, emojis, special character, hastegs from the title and comment column.

Moreover, with the help of the nltk.word\_tokenize() function we split the text into words. aftwe are utilizing stopwords from nltk.corpus, which consists of a catalogue of commonly used stop-words in multiple languages. These stop-words are prevalent words that lack substantial meaning and are typically excluded to concentrate on the critical words within the text. By employing these stop-words, we can effectively filter out irrelevant texts. Furthermore, we are employing the WordNetLemmatizer from nltk.stem, which offers lemmatization functionality. Lemmatization involves the process of reducing words to their fundamental or dictionary form, while ensuring that the word is relevant to the specific language. NLTK's WordNetLemmatizer leverages WordNet, an English lexical database, to retrieve the root form of words, aiding in the normalization of the text and the reduction of inflectional forms to a standardized base form.

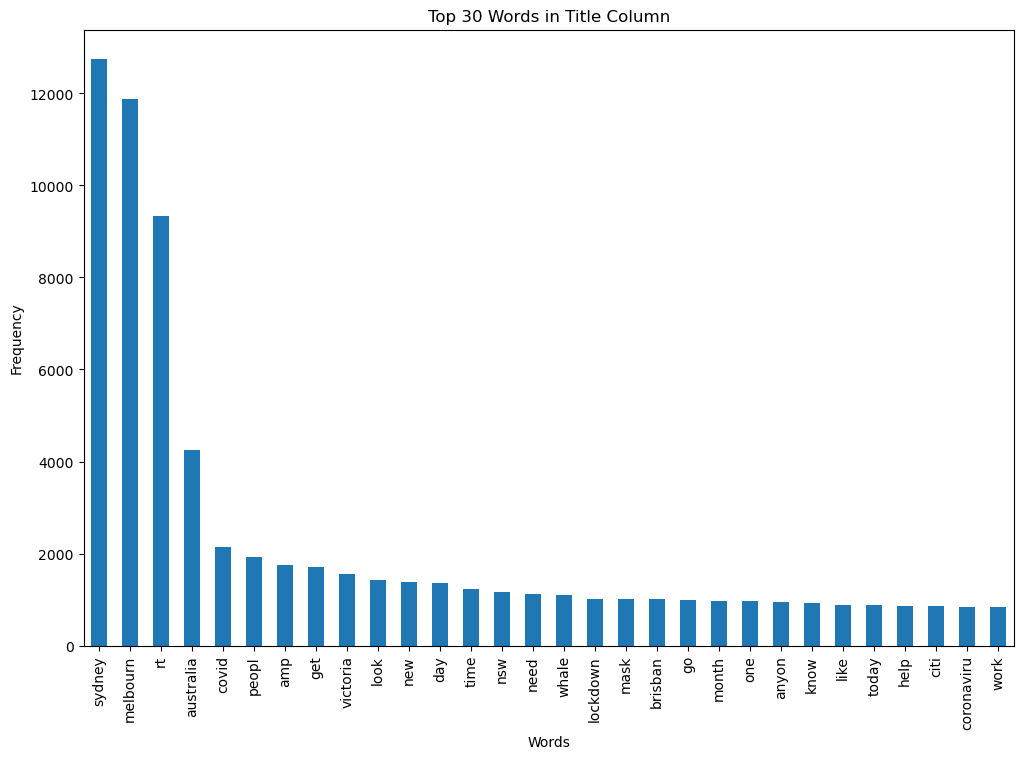
**Initial Explorations**

After preprocessing and cleaning the data, we have a total of 28,790 data in our dataset. Now let’s see some generalized statistics of our data set.

A blue rectangular object with white text

Description automatically generatedWe made a graph of post distribution by location and source. We can see that we almost have similar quantity of data on each segment. So, we can have a more realistic look on the data analysis part.

To quickly scan the most frequent terms, we've extracted the top 30 words. It provides a basic understanding of the type of data we are analysing. Several intriguing words include- Sydney, Melbourne, RT(retweet), australia etc.



**Analysis approach**

After collecting and preprocessing all the required data we decided to follow these approaches for the analysis

*Selection of lexicon*

Using the VADER (Valence Aware Dictionary and sEntiment Reasoner), we aim to find the underlying sentiments within the collected social media data. By employing these lexicons, we can assess the polarity and emotional context of the discussions, providing a comprehensive understanding of the sentiment dynamics prevalent in the online conversations.

*Selection of topic modelling approach*

Utilizing sophisticated techniques such as latent Dirichlet allocation (LDA) implemented through the SKLearn library, we aim to categorize and identify the key areas of interest and conversation, thereby offering a comprehensive picture of the prominent cultural elements that dominate the social media landscapes of these two cities.

*Selection of Graph modelling method*

Using the NetworkX graph library, which was provided within RMIT help lab sessions, we intend to construct and analyse the network structures derived from the social media interactions.

**Network Graph**

In the graph we can see the visualization of the user interaction in Melbourne and Sydney subreddit. If we visualise many users’ interaction in a graph it would be a mess of nodes and A network of colored dots and lines

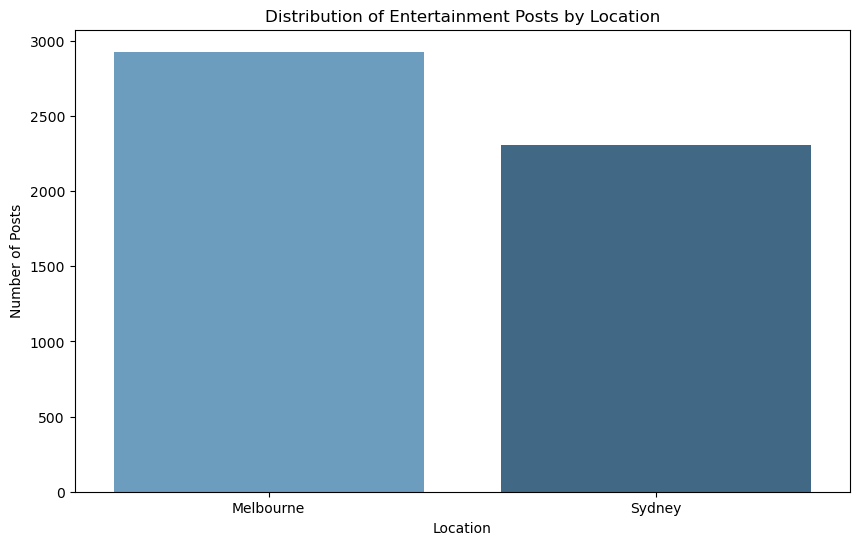
Description automatically generatedA network of colored dots and lines

Description automatically generatededges. So, we decided to Louvain community detection algorithm and visualise it with node clustering. For the user interaction graph we also decided to do it directly using PRAW API instead of the collected data because it will take a lot of computational power and time. We can say that, in terms of user interactions, Melbourne and Sydney subreddit is pretty similar.

**Topic Modelling**

Topic modeling serves as a crucial tool in unraveling the underlying thematic structures within the amassed social media data. By implementing the LDA algorithm through the SKLearn library, we aim to find the prevalent themes and subjects that dominate the online discussions within the Melbourne and Sydney communities.

For the LDA topic modelling we filtered the data using a list of keywords of entertainment and food to segment the data. The entertainment key words are- concert, movie, theatre, park, music, festival, show, performance, entertainment, event, audience, ticket, file, screen, cinema, artists, stage, spectacle, band, recreation, art, gallery, sports, game, and hobby. Here is the distribution of entertainment related post graph-



The food key words are - Food, Restaurant, Dine, Cuisine, Eat, Meal, Dish, Menu, Culinary, Gourmet, Delicacy, Taste, Flavour, Dining, Cooking, Chef, Recipe, Ingredients, Cook, cozy. Here is the distribution of food related post graph-

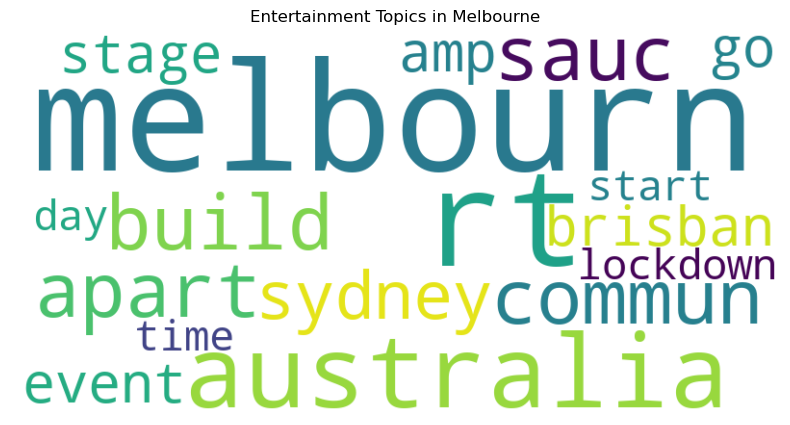
A blue rectangular bars with white text

Description automatically generated

In our data, we can see that in Melbourne people are more likely to talk about entertainment than the people of Sydney. Maybe it is a good indication of Melbourne being the cultural capital of Australia. On the other hand, on the topic of food, people of Sydney are more concerned than people of Melbourne.

The word cloud for Melbourne's entertainment scene highlights key recurring terms such as "event," "stage," "lockdown," and "community," indicating a strong emphasis on community engagement and cultural events despite the challenges posed by lockdowns. Part of our data is from the time frame of covid period, that’s why they are concerned about Additionally, the mention of "apart" and "build" suggests a focus on building and maintaining a sense of togetherness and resilience within the local community, underscoring the city's strong communal spirit. In contrast, the word cloud representing Sydney's entertainment landscape showcases terms like "covid," "park," and "music," indicating a significant focus on outdoor activities and recreational pursuits, particularly in the context of managing the impacts of the COVID-19 pandemic. The inclusion of "ilovesydney" suggests a concerted effort to promote Sydney as a vibrant and resilient city.

A close up of words

Description automatically generated

Examining the word cloud for Melbourne's food culture, prominent terms such as "great meal," "best," and "eat" emphasize the city's thriving culinary scene and the appreciation for diverse dining experiences. The mention of "covid" reflects the resilience of Melbourne's food industry amid the pandemic, underscoring the city's determination to maintain its rich gastronomic culture and provide exceptional dining experiences despite challenging circumstances.

The word cloud for Sydney's food culture showcases terms like "restaurant," "delightful," and "gourmet," reflecting the city's penchant for upscale dining experiences and gourmet delights. The inclusion of "NWS" suggests a strong regional focus, highlighting the city's connection to the broader New South Wales culinary landscape. Additionally, the use of terms like "travel" and "Australia" indicates a thriving culinary tourism scene, underlining Sydney's position as a hub for diverse and exquisite gastronomic experiences that attract both locals and tourists alike.

A close up of words

Description automatically generated

A close up of words

Description automatically generated

**Sentiment analysis**

In the sentiment analysis we will use VADER sentiment analysis to find the people’s view on the entertainment and food scene of Melbourne and Sydney. In the following graph we can see the following findings-

* Melbourne and Sydney have similar sentiment patterns: The graphs for Melbourne and Sydney have similar shapes and peaks. This means that people in both cities have similar opinions and preferences when it comes to entertainment and food. They may share similar tastes, cultures, and experiences.

A group of graphs with different colored lines

Description automatically generated

* Food posts have more variation than entertainment posts: The graphs for food posts have wider curves and lower peaks than the graphs for entertainment posts. This means that people have more diverse and nuanced feelings when they talk about food than when they talk about entertainment. They may have different preferences, expectations, and experiences with food.
* Entertainment posts are more positive than food posts: The graphs for entertainment posts have a higher frequency of positive sentiment than the graphs for food posts. This means that people are more likely to express happiness, satisfaction, or excitement when they talk about entertainment than when they talk about food.
* Melbourne food posts have more variation in the positive and negative area then in the post of Sydney, which is more neutral. People have Melbourne have more passion about food then the people of Sydney.

**Conclusion**

In conclusion, the comparative analysis of cultural disparities in entertainment preferences and online discourse on local food culture between Melbourne and Sydney reveals distinctive inclinations and sentiments within each city. Melbourne showcases a resilient cultural spirit through its emphasis on community engagement and diverse dining experiences, while Sydney highlights its vibrant recreational pursuits and upscale culinary scene, emphasizing regional focus and culinary tourism. Recognizing and appreciating these unique cultural identities within the urban landscape of Australia fosters a deeper understanding of community dynamics, paving the way for more effective communication strategies and promoting a sense of belonging and shared cultural heritage.

**References**

[1] W. Jiang, “Australian Cities - Tweets,” *Kaggle.com*, 2020. Available: https://www.kaggle.com/datasets/wjia26/australian-cities-tweets/data. [Accessed: Oct. 20, 2023]

[2] “NLTK :: nltk.tokenize package,” *Nltk.org*, 2023. Available: https://www.nltk.org/api/nltk.tokenize.html. [Accessed: Oct. 20, 2023]

[3] “NLTK :: nltk.corpus package,” *Nltk.org*, 2023. Available: https://www.nltk.org/api/nltk.corpus.html. [Accessed: Oct. 20, 2023]

[4] “NLTK :: nltk.stem.wordnet,” *Nltk.org*, 2023. Available: https://www.nltk.org/\_modules/nltk/stem/wordnet.html. [Accessed: Oct. 20, 2023]

[5]“Python Sentiment Analysis using VADER,” *GeeksforGeeks*, Jan. 23, 2019. Available: https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/. [Accessed: Oct. 20, 2023]

[6] N. Seth, “Part 2: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn,” *Analytics Vidhya*, Jun. 28, 2021. Available: https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/. [Accessed: Oct. 20, 2023]

[7] “NetworkX — NetworkX documentation,” *Networkx.org*, 2023. Available: https://networkx.org/. [Accessed: Oct. 20, 2023]

‌[8] “Louvain - Neo4j Graph Data Science,” *Neo4j Graph Data Platform*, 2023. Available: https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/. [Accessed: Oct. 20, 2023]

‌

‌

‌

‌