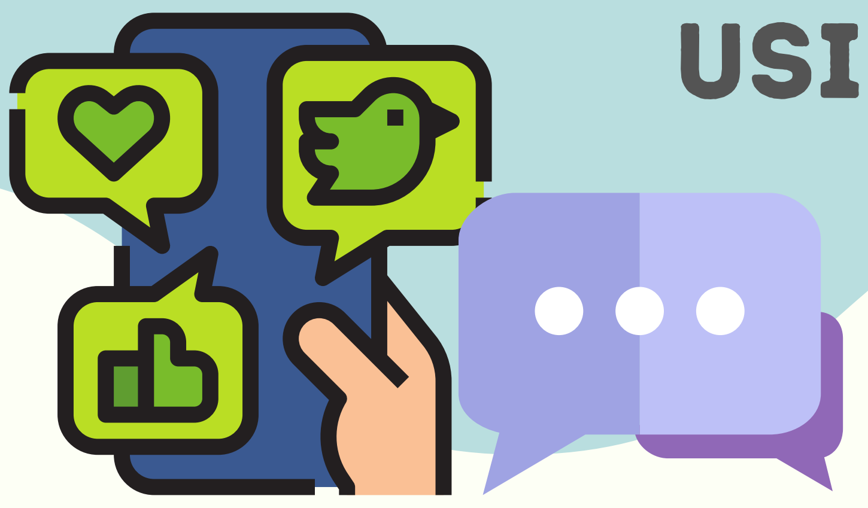
OPINION ANALYSIS ON TWEETS ABOUT HOUSING

USING NATURAL LANGUAGE PROCESSING TECHNIQUE



PLEASE FEEL FREE TO GIVE ADVICE FOR IMPROVEMENT

BY THAPANEE PUNNAVIRAT
thapanee.punnavirat@gmail.com

INTRODUCTION

Housing crisis becomes a topic creating concerns for many people around the world mostly due to a lack of housing that is affordable to purchase or rent. Referring to the World Bank (2021)'s statement, 1.6 billion people tend to be impacted by the global housing shortage by 2025. In many countries, house price increases dramatically without corresponding to income growth which accelerates at a slower rate. According to the data of Demographia International Housing Affordability updated on March 2022 by Urban Reform Institute and Frontier Centre for Public Policy, the cities such as Hong Kong, Sydney, and Dublin have ratings on unaffordability at a severe level (median multiple at 5.1 or above).

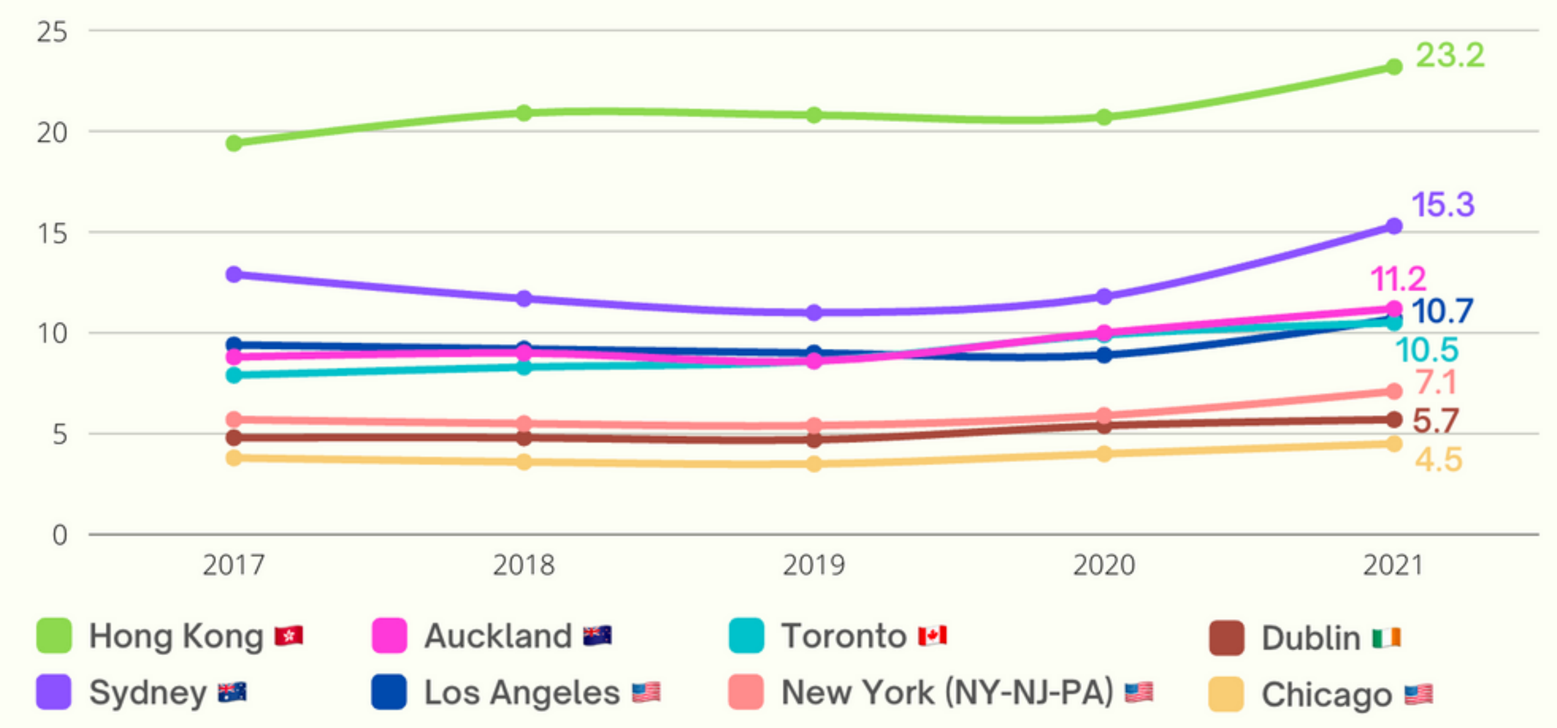
The housing crisis has been caused by many factors. One of the factors is the impact of low-interest rates and the shift to remote work in many countries during the COVID-19 pandemic, which accelerates the demand for houses. However, many aspects of housing problems are still implicit and not yet mentioned widely.

Learning opinions relating to housing from society in a most recent and constant manner will lead to new insights into housing issues. For these reasons, this project has an objective to study the top-mentioned keywords, the sentiments, and the summaries of the tweets about housing by using the Natural Language Processing technique, which some challenges encountered from the sentiment analysis will also be introduced.

HOUSING AFFORDABILITY (GLOBAL CITIES)

*the ratio of average house prices to average income

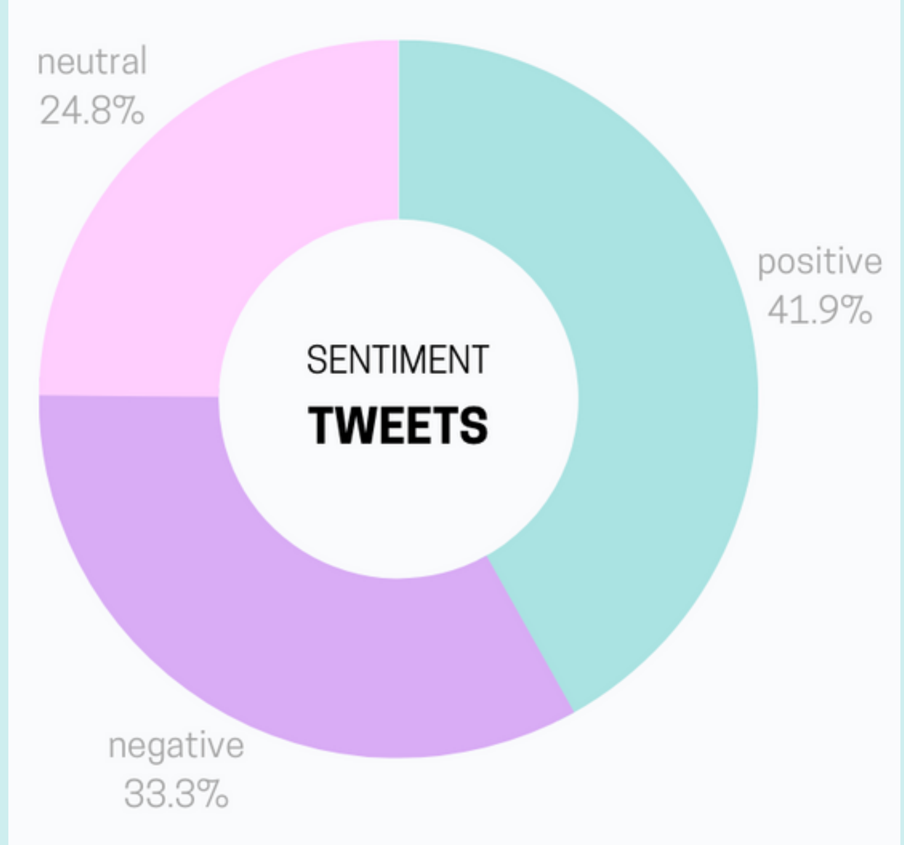
Rating	Median Multiple
Severely unaffordable	5.1 & over
Seriously unaffordable	4.1 - 5.0
Moderately unaffordable	3.1 - 4.0
Affordable	3.0 or less

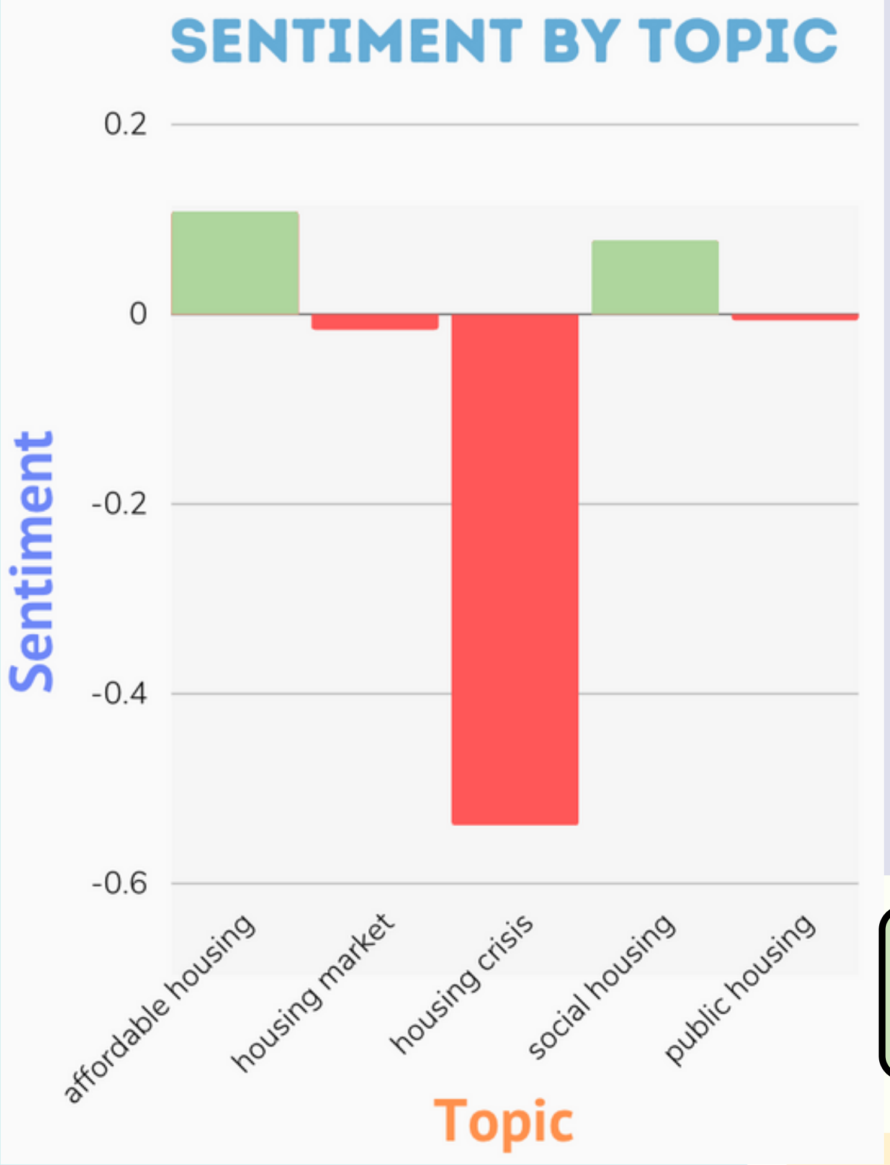


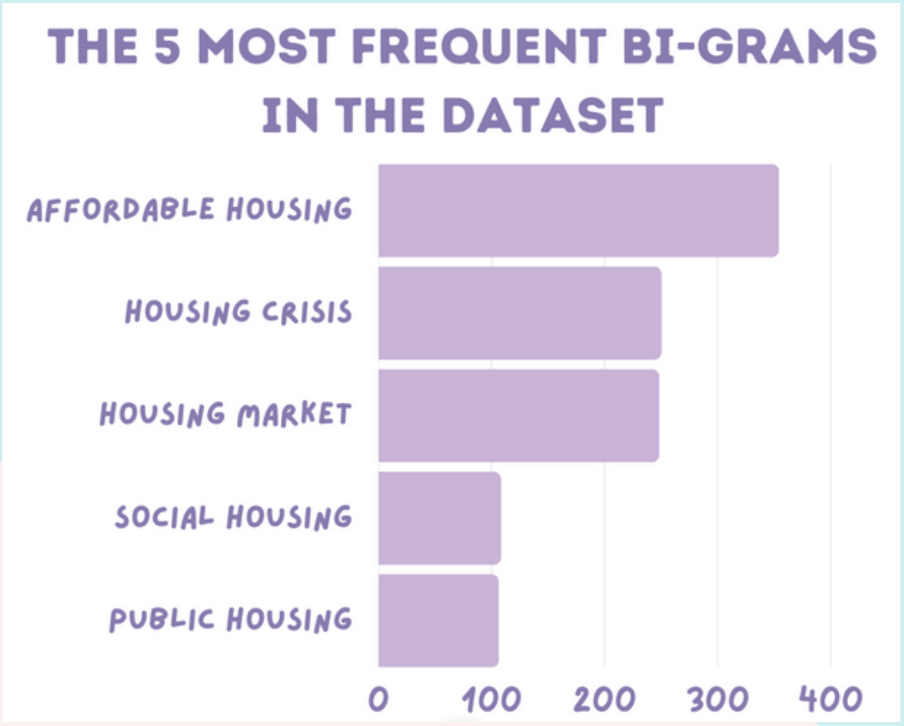
Source: Demographia International Affordability Survey

RESULTS

After extracting words by using the Bi-gram technique, "affordable housing", "housing market", "housing crisis", "social housing", and "public housing" are ranked the top 5 mentioned keywords in tweets about housing respectively. When analyzing the sentiment of tweets based on these top keywords, "affordable housing" and "social housing" reflect positive sentiment, while the tweets with the "housing market", "housing crisis", and "public housing" keywords are more negative side. However, the overall outcome of this sentiment analysis shows that the majority of the tweets have positive sentiment (41.9%), followed by negative sentiment at 33.3% and neutral sentiment at 24.8%.

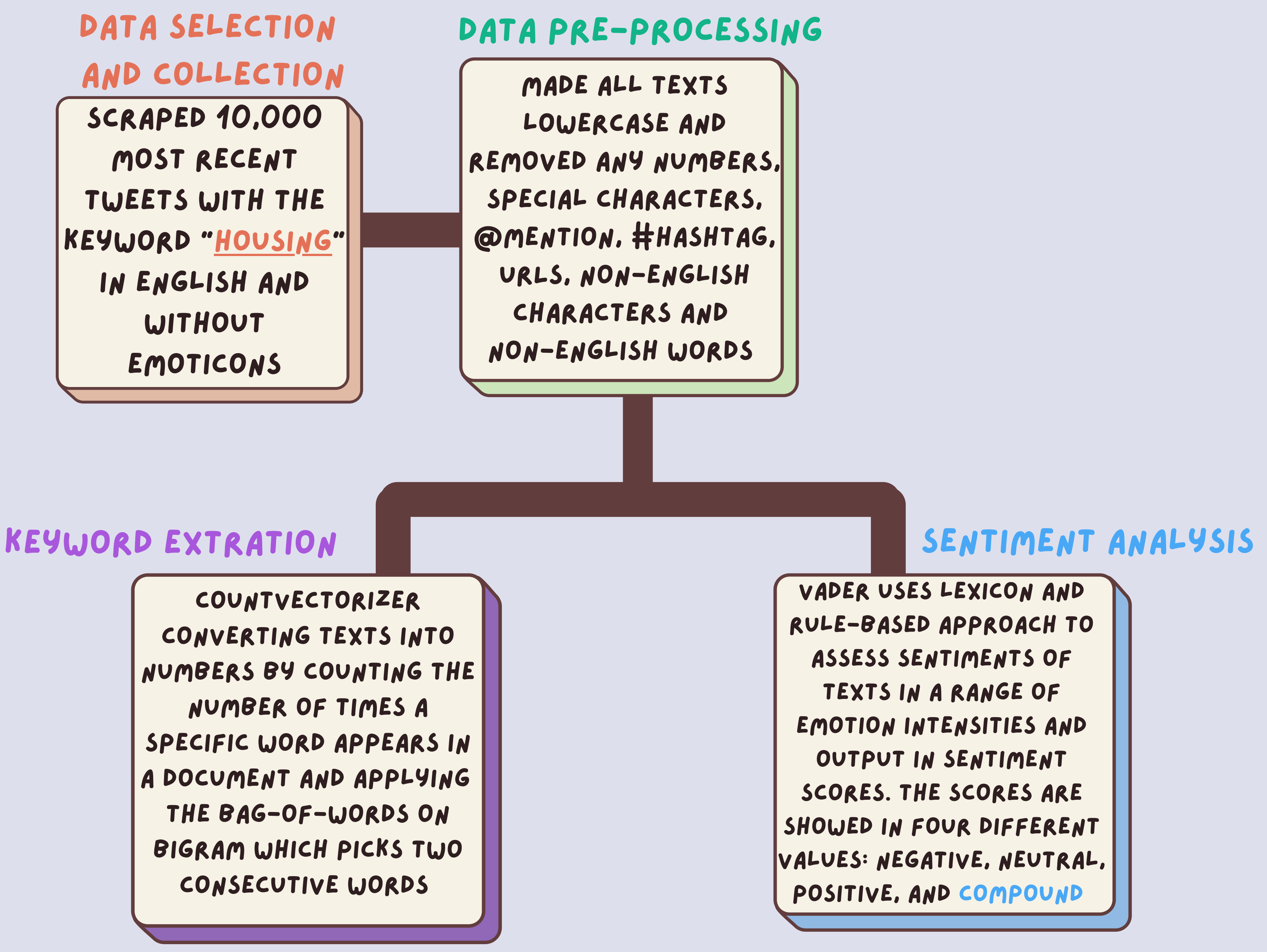






REFERENCES:
Calderon, P. (2017). VADER Sentiment Analysis Explained.
Delteil, C. (n.d.). Unsupervised Sentiment Analysis With Real-World Data: 500,000 Tweets on Elon Musk.
Educative. (2023). CountVectorizer in Python.
Jain, P. (2021). Basics of CountVectorizer.
The Housing Agency. (2022). House Price to Income Ratio.
World Bank. (2021). Can new technology solve the global housing crisis?

WORKFLOW



DATA SELECTION AND COLLECTION

SCRAPED 10,000 MOST RECENT TWEETS WITH THE KEYWORD "HOUSING" IN ENGLISH AND WITHOUT EMOTICONS

DATA PRE-PROCESSING

MADE ALL TEXTS LOWERCASE AND REMOVED ANY NUMBERS, SPECIAL CHARACTERS, @MENTION, #HASHTAG, URLS, NON-ENGLISH CHARACTERS AND NON-ENGLISH WORDS

KEYWORD EXTRATION

COUNTVECTORIZER CONVERTING TEXTS INTO NUMBERS BY COUNTING THE NUMBER OF TIMES A SPECIFIC WORD APPEARS IN A DOCUMENT AND APPLYING THE BAG-OF-WORDS ON BIGRAM WHICH PICKS TWO CONSECUTIVE WORDS

SENTIMENT ANALYSIS

VADER USES LEXICON AND RULE-BASED APPROACH TO ASSESS SENTIMENTS OF TEXTS IN A RANGE OF EMOTION INTENSITIES AND OUTPUT IN SENTIMENT SCORES. THE SCORES ARE SHOWED IN FOUR DIFFERENT VALUES: NEGATIVE, NEUTRAL, POSITIVE, AND COMPOUND

HOW COUNTVECTOR WORKS?

COUNTVECTORIZER CONVERTS ALL WORDS TO LOWERCASE AND REMOVES ANY SPECIAL CHARACTERS. THEN, IT USES THE BAG-OF-WORDS TECHNIQUE TO TRANSFORM TEXT DATA OR WORDS TO NUMERICAL VALUES BY COUNTING HOW OFTEN THE SPECIFIC WORD APPEARS IN A SENTENCE OR A DOCUMENT.

Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']

↓

	The	quick	brown	fox	jumps	over	lazy	dog
Data	2	1	1	1	1	1	1	1

SOURCE: EDUCATIVE, 2023

COMPOUND SCORE <= -0.05 NEGATIVE
COMPOUND SCORE >= 0.05 POSITIVE
COMPOUND SCORE > -0.05 NEUTRAL
< 0.05

HOW VADER WORKS?

VADER SCANS THE TEXT FOR KNOWN SENTIMENTAL FEATURES BY COMPARING TO THE AVAILABLE LEXICAL DICTIONARY, MODIFIED THE INTENSITY AND POLARITY BASED ON THE GRAMMATICAL RULES (PUNCTUATION, CAPITALIZATION, ADVERBS AND CONTRASTIVE CONJUNCTIONS), SUMMED UP THE SCORES OF FEATURES FOUND WITHIN THE TEXT AND NORMALIZED THE FINAL SCORE TO (-1, 1) USING FUNCTION:

$$\frac{x}{\sqrt{x^2+\alpha}}$$

ALPHA IS SET TO BE 15, APPROXIMATING THE MAXIMUM EXPECTED VALUE OF X.

CHALLENGES

DOMAIN-SPECIFIC DICTIONARY OR CORPUS

The polarity of opinion words may vary for each specific domain. A positive sentence in one domain may be a negative sentence in another domain. For instance, the word "huge". "This chair is huge" may sound positive, but "This card is too huge to put in my bag." may sound negative.

SARCASM AND IRONIC PHRASES

Sometimes, the model cannot detect that the sentences containing positive words are actually sarcastic. For example, the model can predict a positive sentiment for the sentence "Brilliant! I am fired today." because it contains the word "Brilliant!"

NEOLOGISM

Languages change all the time. New words are created by groups of people from time to time. The lexical dictionary or corpus needs to be updated constantly to be able to predict sentiment analysis accurately.