MCST 1043
RESEARCH DESIGN AND ANALYSIS

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SENTIMENT ANALYSIS ON HOTEL REVIEW USING MACHINE LEARNING

CANDIDATE : NURFATINI ATIQAH BINTI HAMIDI
LECTURER : ASSOC PROF DR MOHD SHAHIZAN OTHMAN
DATE : 17 JANUARY 2025

FACULTY OF COMPUTING,
UNIVERSITI TEKNOLOGI MALAYSIA
www.utm.my
innovative • entrepreneurial • global

univteknologimalaysia utm_my utmofficial

Content

01 **Introduction**

02 **Literature Review**

03 **Research Methodology**

04 **Initial Findings**

05 **Conclusion**

INTRODUCTION

INTRODUCTION

PROBLEM BACKGROUND AND STATEMENT

Problem Background:

- Online reviews significantly impact consumer decision-making, particularly in the hospitality industry
- Platforms like TripAdvisor, Agoda, and Booking.com receive thousands of reviews reflecting customer experiences and emotions
- Manual tracking and extracting information from extensive reviews is time-consuming and inefficient

Problem Statement:

- Inconsistencies between written reviews and star ratings lead to confusion and misinterpretation.
- Reviews are critical for shaping a hotel's online reputation and influencing customer decisions

INTRODUCTION

OBJECTIVES

01

To conduct exploratory data analysis to identify patterns of hotel reviews

02

To design and implement sentiment analysis that predict the review either positive or negative.

03

To conduct comprehensive evaluations on the develops predictive model and build an interactive dashboard.

INTRODUCTION

RESEARCH SCOPE

Open-source dataset from Kaggle

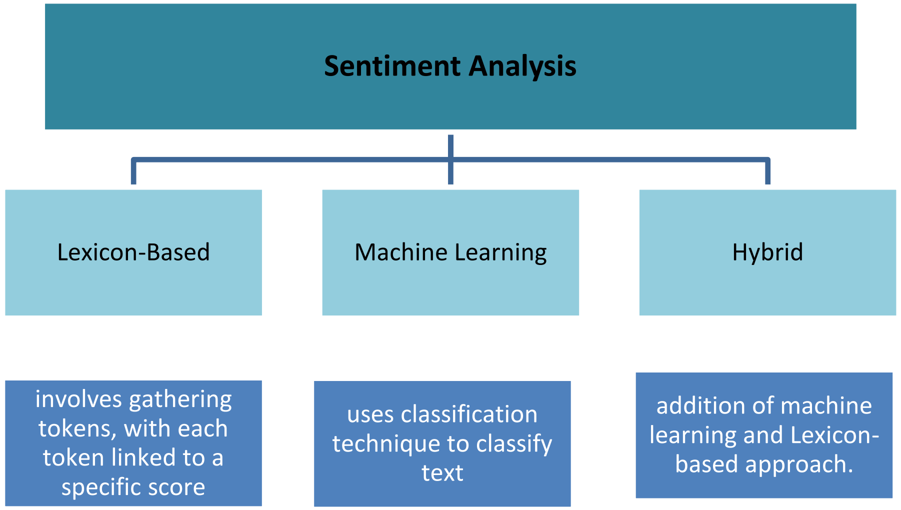
Two Sentiment Analysis (+ve and –ve)

Machine Learning: TF-IDF and K-Means

LITERATURE REVIEW

LITERATURE REVIEW

THEORETICAL FRAMEWORK



Sentiment Analysis

Lexicon-Based

Machine Learning

Hybrid

involves gathering tokens, with each token linked to a specific score

uses classification technique to classify text

addition of machine learning and Lexicon-based approach.

Type of approach	Advantages	Limitation
Machine Learning	The capability to adjust and develop trained models for particular uses and situations.	The limited applicability to new data arises from the need for labelled data, which can be expensive or even unaffordable.
Lexicon-based	Broader term inclusion, annotated data, and the process of learning are not necessary.	A restricted set of words in the lexicons is assigned a particular sentiment orientation and score for each word
Hybrid	Lexicon/learning symbiosis involves identifying and gauging sentiment at the conceptual level, along with reduced sensitivity to shifts in topic domain.	Noisy reviews

LITERATURE REVIEW

MACHINE LEARNING APPROACH

The Term Frequency-Inverse Document Frequency (TF-IDF)

a prevalent technique used to assess the significance of a word within a document

Random Forest

Random forest is an ensemble technique that generates multiple decision trees, which are then combined into a forest. Each tree in the Random Forest makes a class prediction and the Random Forest determines the outcome based on tge majority of votes.

Support Vector Machine (SVM)

SVM finds the optimal hyperplane (a decision boundary) that separates data points into distinct classes with the maximum margin between them.

LITERATURE REVIEW

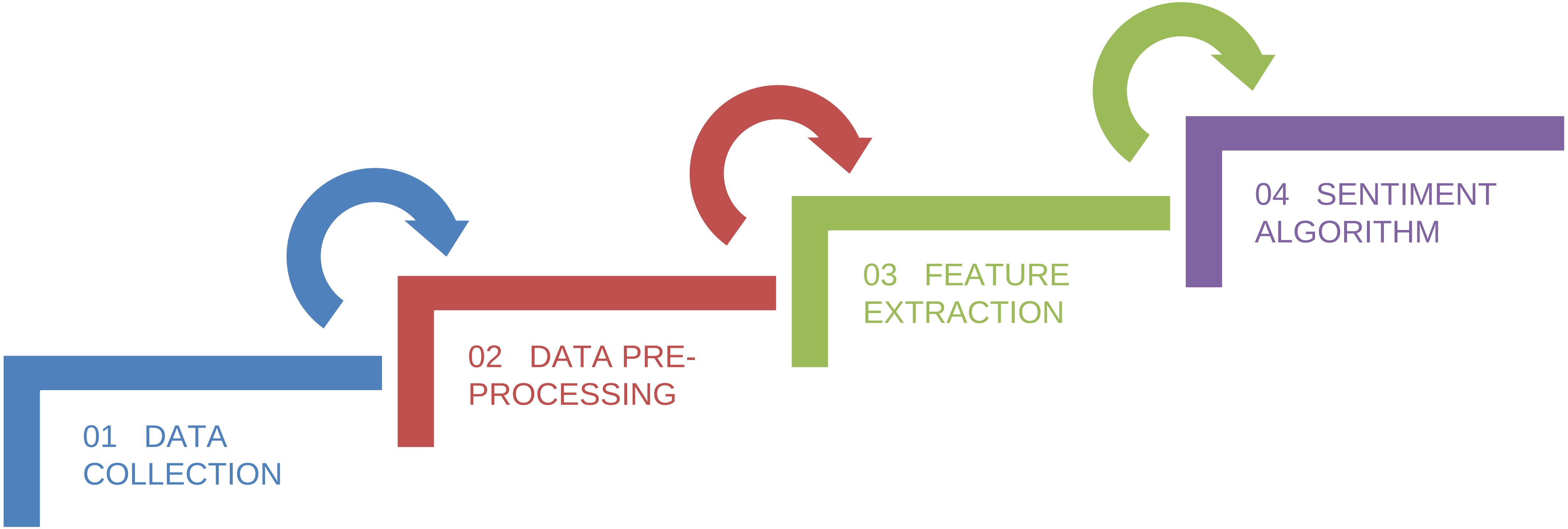
DATA AND METHODS COMPARISON

Author	Dataset	Features	Classifier	Accuracy (%)
(Rathi et al., 2018)	Sentiment140, Polarity Dataset, and University of Michigan dataset	TF-IDF	AdaBoost	67
(Makhmudah et al., 2019)	Tweets related to homosexuals	TF-IDF	Support Vector Machine (SVM)	99.5
(Gupta et al., 2019)	Sentiment140	TF-IDF	Neural Network	80
(Alsalman, 2020)	Arabic Tweets	TF-IDF	Multinomial Naïve Bayes	87.5
(Alzyout et al., 2021)	Self-collected dataset	TF-IDF	Support Vector Machine (SVM)	78.25

RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

RESEARCH FRAMEWORK



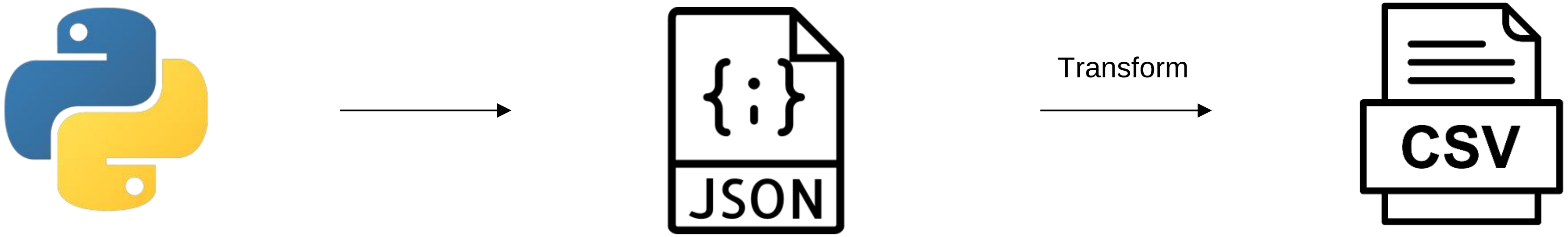
01 DATA COLLECTION

02 DATA PRE-PROCESSING

03 FEATURE EXTRACTION

04 SENTIMENT ALGORITHM

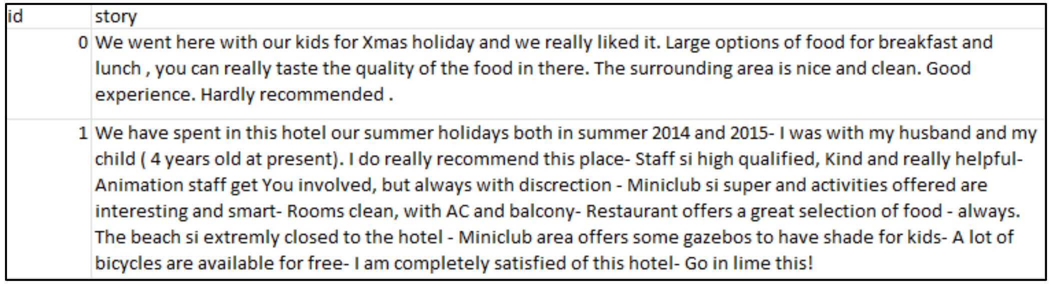
DATA COLLECTION



Size:
1,100,001

Attribute:

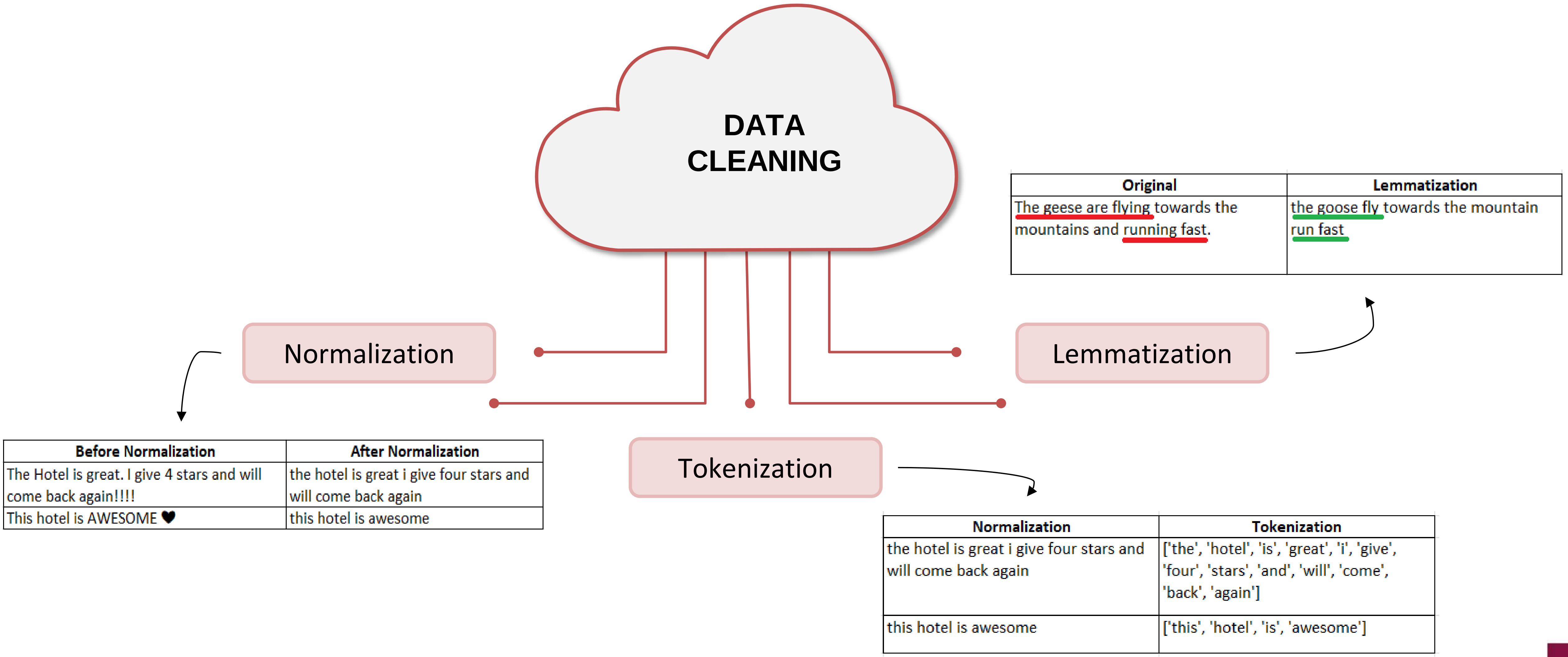
- Id
- story



id	story
0	We went here with our kids for Xmas holiday and we really liked it. Large options of food for breakfast and lunch , you can really taste the quality of the food in there. The surrounding area is nice and clean. Good experience. Hardly recommended .
1	We have spent in this hotel our summer holidays both in summer 2014 and 2015- I was with my husband and my child (4 years old at present). I do really recommend this place- Staff si high qualified, Kind and really helpful- Animation staff get You involved, but always with discrection - Miniclub si super and activities offered are interesting and smart- Rooms clean, with AC and balcony- Restaurant offers a great selection of food - always. The beach si extremly closed to the hotel - Miniclub area offers some gazebos to have shade for kids- A lot of bicycles are available for free- I am completely satisfied of this hotel- Go in lime this!

RESEARCH METHODOLOGY

DATA PRE-PROCESSING



Original	Lemmatization
The geese are flying towards the mountains and running fast.	the goose fly towards the mountain run fast

Before Normalization	After Normalization
The Hotel is great. I give 4 stars and will come back again!!!!	the hotel is great i give four stars and will come back again
This hotel is AWESOME ♥	this hotel is awesome

Normalization	Tokenization
the hotel is great i give four stars and will come back again	['the', 'hotel', 'is', 'great', 'i', 'give', 'four', 'stars', 'and', 'will', 'come', 'back', 'again']
this hotel is awesome	['this', 'hotel', 'is', 'awesome']

TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF)

- TF-IDF highlights words that are frequent in a specific document but rare across the entire corpus, making them more relevant for understanding the document's content.

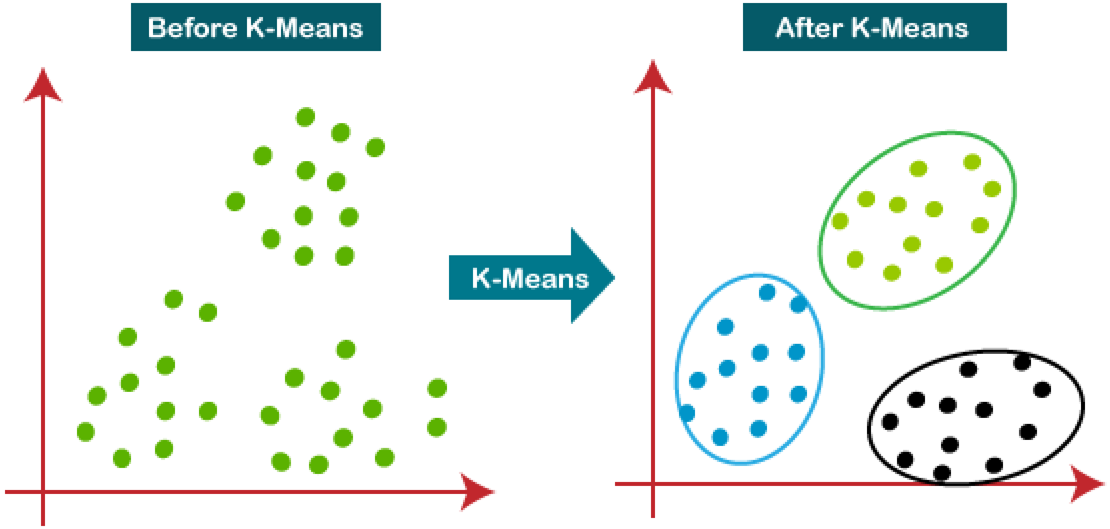
$$W_{ij= tf_{ij}} \log(\frac{D}{df_j})$$

EXAMPLE:

consider a document that consists of 100 words, with the word “happy” occurring 10 times. In this case, the term frequency would be calculated as 10/100=0.1. Now, let's assume there are 50000 documents in total, and only 500 of those contain the word “happy.” Therefore, the IDF (happy) can be expressed as 50000/500=100, resulting in log(100) = 2. Consequently, the TF-IDF (happy) would be 0.1*2= 0.2

K-MEANS CLUSTERING

- K-Means clustering is one of the machine learning algorithms used for grouping data points into clusters based on their similarity

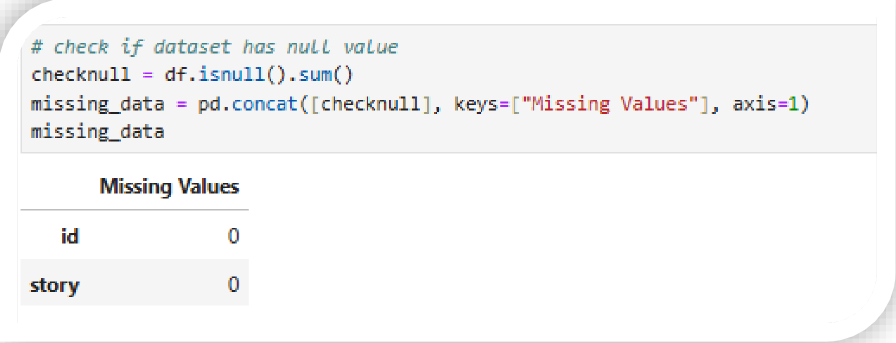


INITIAL FINDINGS

INITIAL FINDINGS

EXPLORATORY DATA ANALYSIS (EDA) python

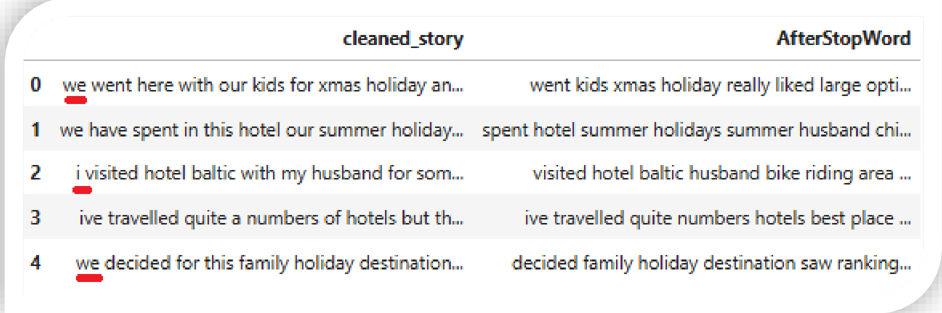
Check missing values



```
# check if dataset has null value
checknull = df.isnull().sum()
missing_data = pd.concat([checknull], keys=["Missing Values"], axis=1)
missing_data
```

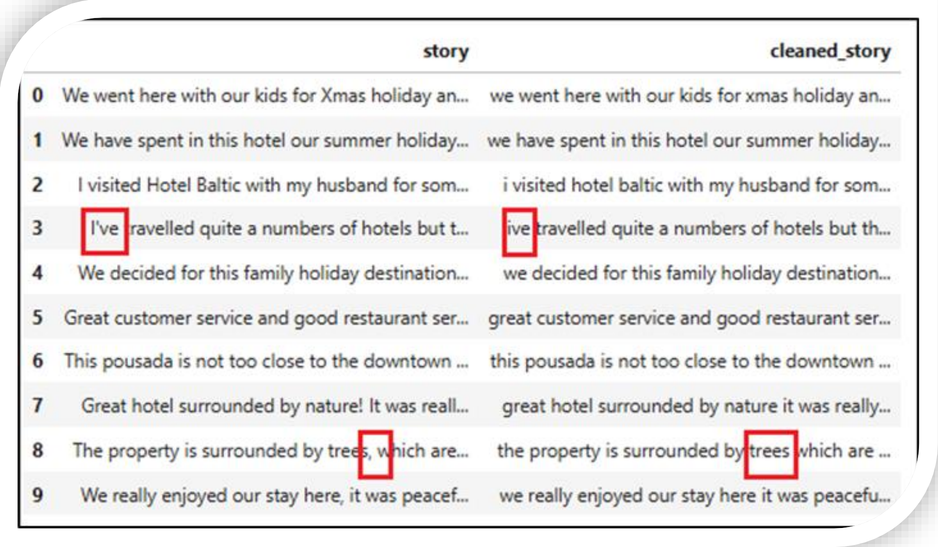
	Missing Values
id	0
story	0

Implement stop words



	cleaned_story	AfterStopWord
0	we went here with our kids for xmas holiday an...	went kids xmas holiday really liked large opti...
1	we have spent in this hotel our summer holiday...	spent hotel summer holidays summer husband chi...
2	i visited hotel baltic with my husband for som...	visited hotel baltic husband bike riding area ...
3	ive travelled quite a numbers of hotels but th...	ive travelled quite numbers hotels best place ...
4	we decided for this family holiday destination...	decided family holiday destination saw ranking...

Remove punctuation

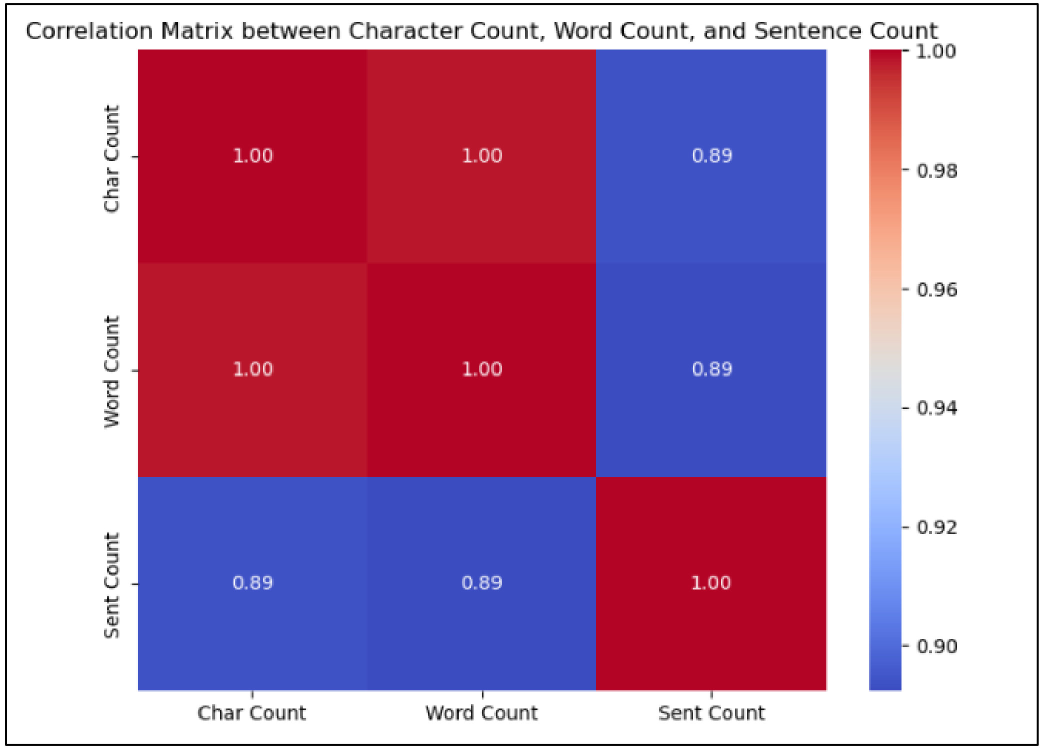


	story	cleaned_story
0	We went here with our kids for Xmas holiday an...	we went here with our kids for xmas holiday an...
1	We have spent in this hotel our summer holiday...	we have spent in this hotel our summer holiday...
2	I visited Hotel Baltic with my husband for som...	i visited hotel baltic with my husband for som...
3	I've travelled quite a numbers of hotels but t...	ive travelled quite a numbers of hotels but th...
4	We decided for this family holiday destination...	we decided for this family holiday destination...
5	Great customer service and good restaurant ser...	great customer service and good restaurant ser...
6	This pousada is not too close to the downtown ...	this pousada is not too close to the downtown ...
7	Great hotel surrounded by nature! It was reall...	great hotel surrounded by nature it was really...
8	The property is surrounded by trees, which are...	the property is surrounded by trees which are ...
9	We really enjoyed our stay here, it was peacef...	we really enjoyed our stay here it was peacefu...

UTM UNIVERSITI TEKNOLOGI MALAYSIA

INITIAL FINDINGS

EXPLORATORY DATA ANALYSIS (EDA) python

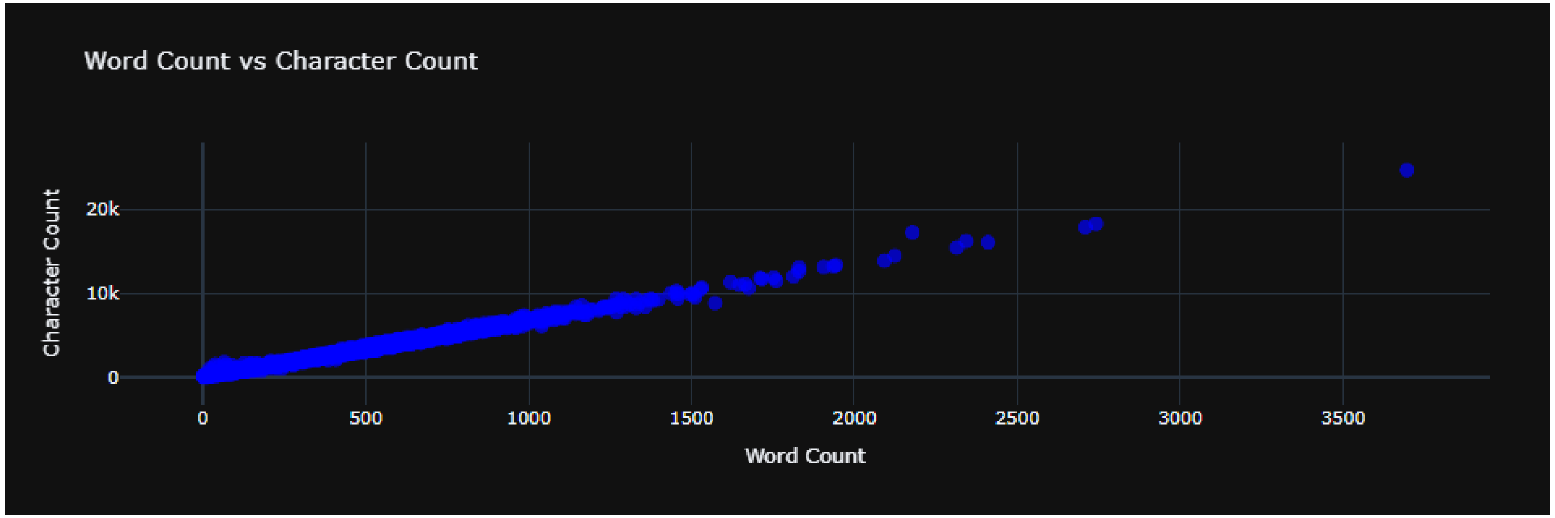


The character count and word count strongly correlate positively. While character count and sentence count have a moderate correlation that depends on sentence length.

INITIAL FINDINGS

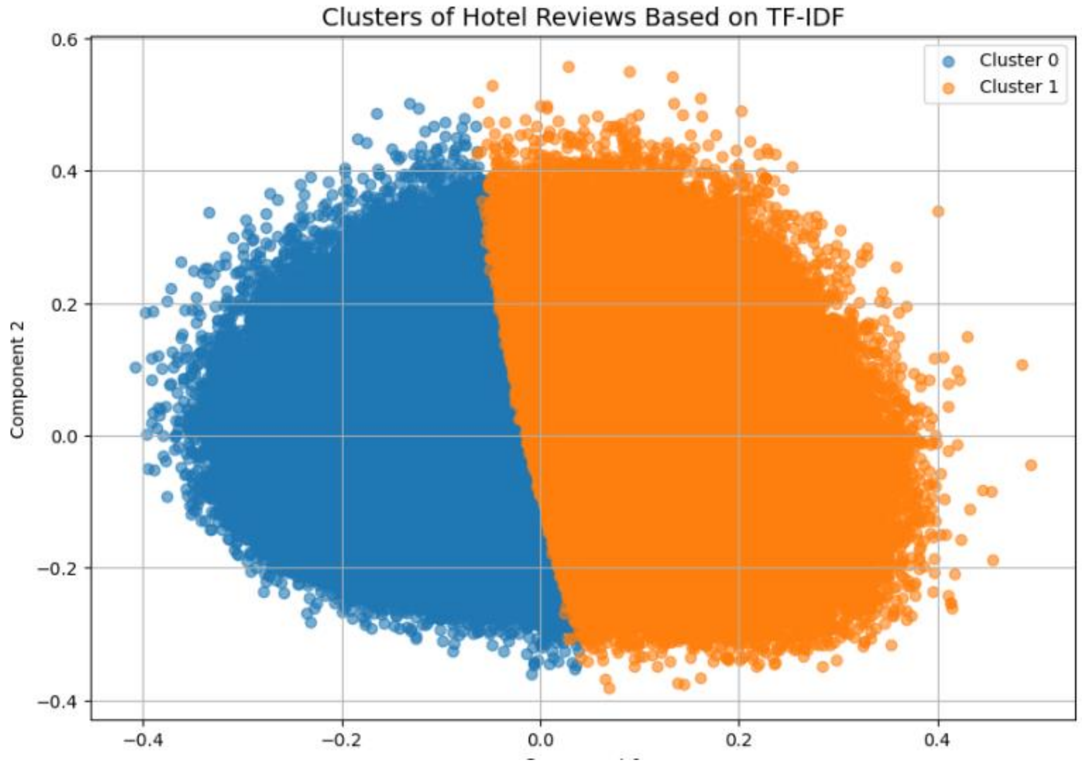
EXPLORATORY DATA ANALYSIS (EDA) python

Positive correlation between word count and character count



INITIAL FINDINGS

EXPLORATORY DATA ANALYSIS (EDA) python



The proportion of two distinct groups, represented by the blue and orange colors. The blue represents the cluster 0 while the orange represents the cluster 1. In this cluster, 0 represent positive and 1 represent negative.

CONCLUSION

CONCLUSION

SUMMARY

This study has achieved the objective, to identify the patterns in hotel reviews:

- From the analysis, it can be found that word count and character count increase together.Word count links better with character count than sentence count (due to cleaning).
- Used TF-IDF to find the 500 most important words.
- Used K-Means to sort reviews into positive and negative groups.
- Sentiments form clear clusters, but some overlap and outliers exist.

CONCLUSION

FUTURE WORKS

Improve Preprocessing:
Use advanced techniques like stemming, lemmatization, and part-of-speech tagging to clean noisy data better.

Expand Sentiment Categories:
Move beyond just positive and negative sentiments. Include neutral or mixed opinions for more detailed analysis.

Use Advanced Models:
Replace traditional methods like K-Means with deep learning models like LSTM or BERT.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

univteknologimalaysia utm_my utmofficial

Thank You

www.utm.my

innovative • entrepreneurial • global