Capstone: Malay Language Sentiment Analysis

Ezra Calis
GA DSI 38

Table of Contents

01

Background

02

Problem Statement

03

Methodology

04

EDA

05

Model Analysis

06

Model Deployment

07

Conclusion

01

BACKGROUND

Sentiment Analysis

What is Sentiment Analysis?

Sentiment analysis allows organisations and individuals to understand views on their actions, and themselves.

Why use Sentiment Analysis?

For organisations and individuals who want to track public opinion on them (ie. reputation management), sentiment analysis is vital to help them filter through enormous amounts of unstructured information.

Where to Get Sentiments?

A key source of sentiments can be obtained from social media.

Problems with Malay Language NLP

Lack of Projects

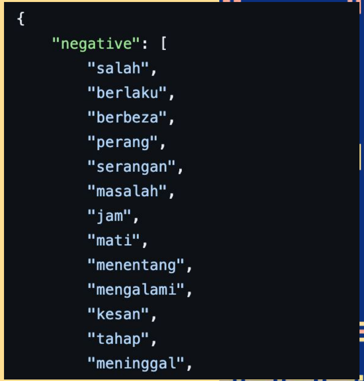
While NLP projects (including sentiment analysis) in major languages (eg. English, Mandarin, French) are a dime a dozen, it has been really scarce for languages like Malay

Malay NLTK

There is only one well-known Natural Language Toolkit library for Malay (**‘Malaya Model’**)

Sentiment analysis is lexicon-based

Quite detailed, but is not representative of social-media slang.



```
{
    "negative": [
        "salah",
        "berlaku",
        "berbeza",
        "perang",
        "serangan",
        "masalah",
        "jam",
        "mati",
        "menentang",
        "mengalami",
        "kesan",
        "tahap",
        "meninggal",
```

Malay Language Social Media

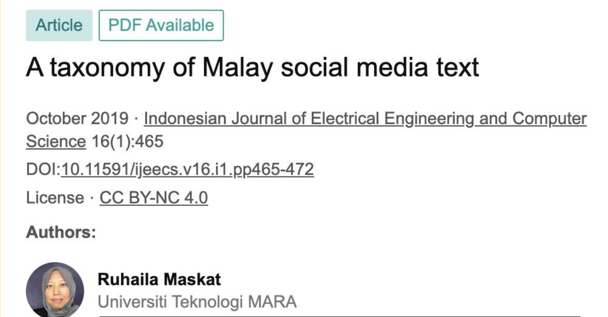
Performing text analytics on Malay social media text is a challenge.

Malay language used in social media differs due to:

1. spelling variations (faham → fhm, phm)
2. Malay-English mix sentence (‘aku suka reaction dia’)
3. slang-based words (abai → buat dek)
4. vowelless words (jangan → jgn)
5. number suffixes (buku-buku → buku2)

Existing lexicon-based sentiment analysis models **do not generalise well to social media comments.**

Companies / organisations looking to **understand sentiments** about them in Malay social media **do not have a good solution.**



Article PDF Available

A taxonomy of Malay social media text

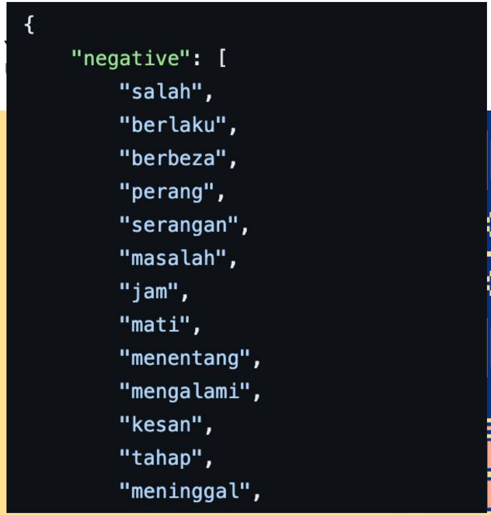
October 2019 · Indonesian Journal of Electrical Engineering and Computer Science 16(1):465

DOI:10.11591/ijeecs.v16.i1.pp465-472

License · CC BY-NC 4.0

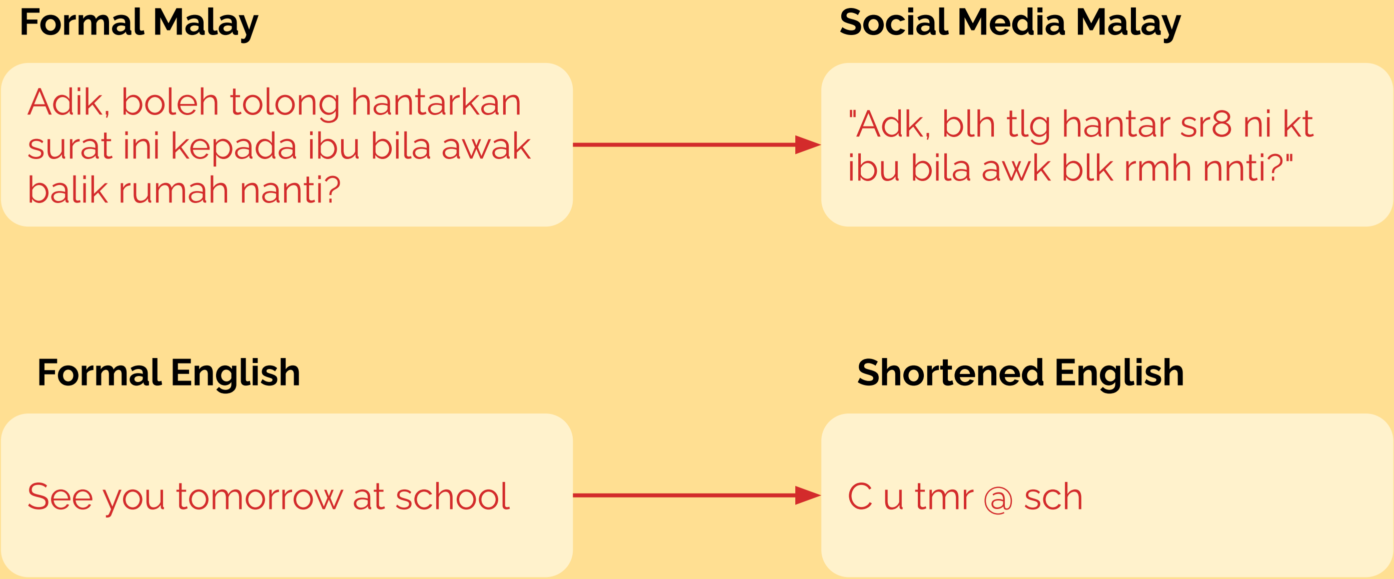
Authors:

Ruhaila Maskat
Universiti Teknologi MARA



```
{
    "negative": [
        "salah",
        "berlaku",
        "berbeza",
        "perang",
        "serangan",
        "masalah",
        "jam",
        "mati",
        "menentang",
        "mengalami",
        "kesan",
        "tahap",
        "meninggal",
```

Example



Formal Malay

Adik, boleh tolong hantarkan surat ini kepada ibu bila awak balik rumah nanti?

Social Media Malay

"Adk, blh tlg hantar sr8 ni kt ibu bila awk blk rmh nnti?"

Formal English

See you tomorrow at school

Shortened English

C u tmr @ sch

02

PROBLEM STATEMENT

PROBLEM STATEMENT

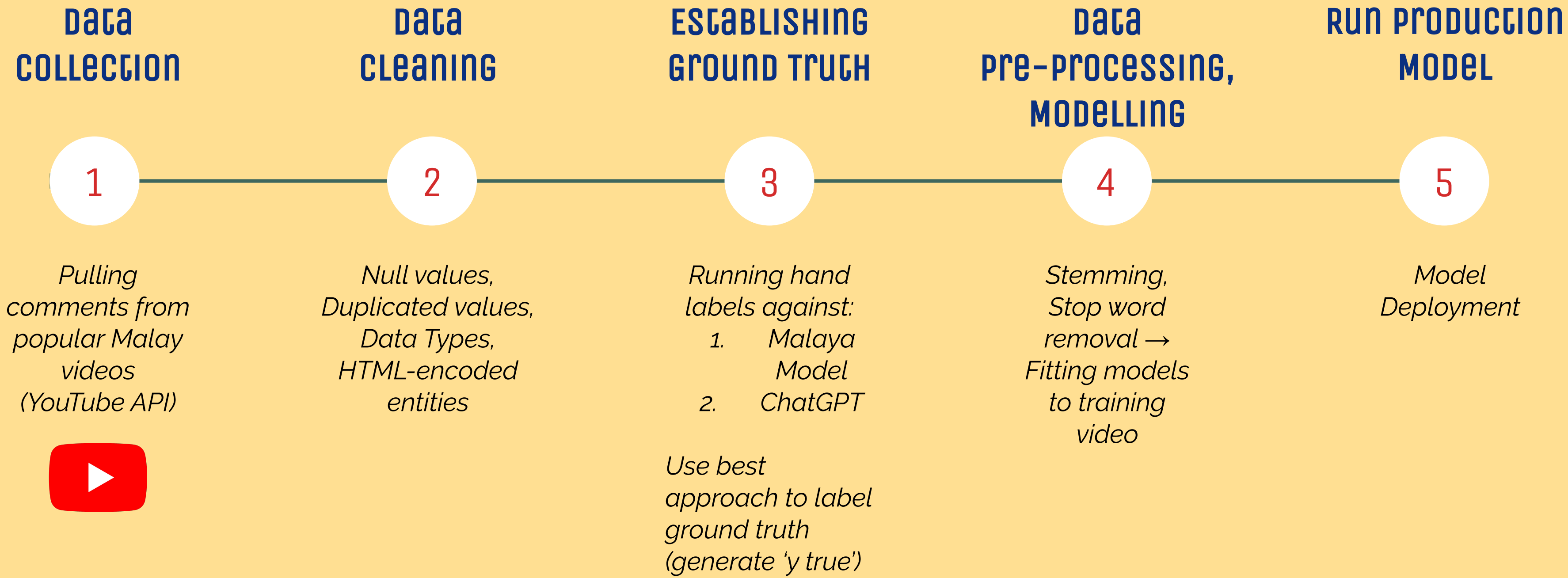
Can we create a **best-in-class sentiment analysis model** for **Malay social media texts**?

Context:
It is currently difficult for organisations / companies to measure social media sentiments in Malay, making reputation management challenging for those who interact with a primarily Malay-speaking audience.

03

METHODOLOGY

METHODOLOGY



Data Collection

1

Pulling comments from popular Malay videos (YouTube API)

Data Cleaning

2

Null values, Duplicated values, Data Types, HTML-encoded entities

Establishing Ground Truth

3

Running hand labels against:

1. *Malaya Model*
2. *ChatGPT*

Use best approach to label ground truth (generate 'y true')

Data Pre-processing, Modelling

4

Stemming, Stop word removal → Fitting models to training video

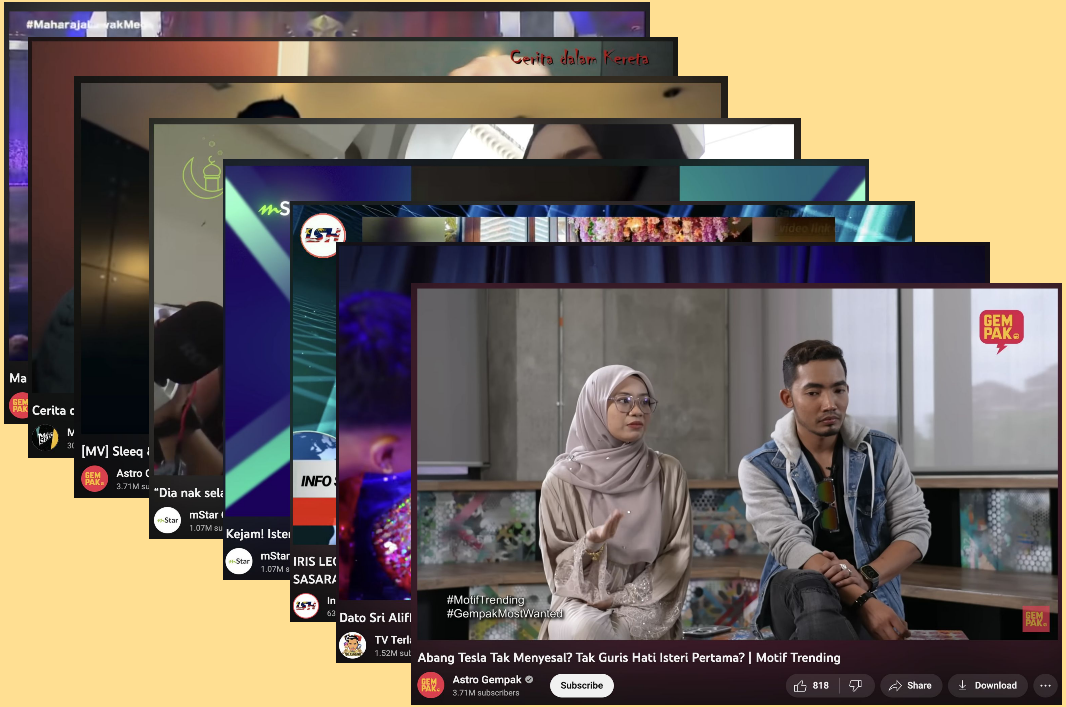
Run Production Model

5

Model Deployment

Data collection

- 9 popular Malay videos
- From 2011 to 2023
- Different types:
  - Positive and negative reactions
  - News, music videos, shows etc.
- **72 million** views total
- **25,000** comments



Data Cleaning

- Null values
- Duplicated values
- Data Types
- HTML-encoded entities

Establishing Ground Truth

1. Manually label sentiments of 100 comments
2. Label sentiments of comments using Malaya model and ChatGPT
3. Compare accuracy score of Malaya model vs ChatGPT
4. Use the more accurate method to label all 25,000 rows

Accuracy score of Malaya model: **0.47**
Accuracy score of ChatGPT: **0.76**

Pre-processing, modelling

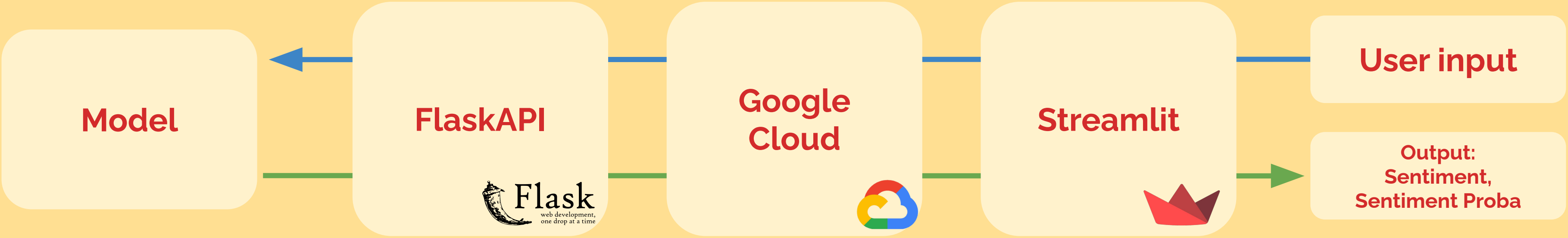
Stemming
pySastrawi (Bahasa Indonesia). Similar word structure

Stopword Removal
Stopwords ISO (collection of stopwords for multiple languages)

Vectorisation
TF-IDF Vectoriser (lower dimensionality, down-weights common terms)

SMOTE / Oversampling / Undersampling
3 methods used to address imbalanced classes, to use the best performing one

Model Deployment


Model
FlaskAPI
Flask
web development, one drop at a time
Google Cloud
Streamlit
User input
Output: Sentiment, Sentiment Proba

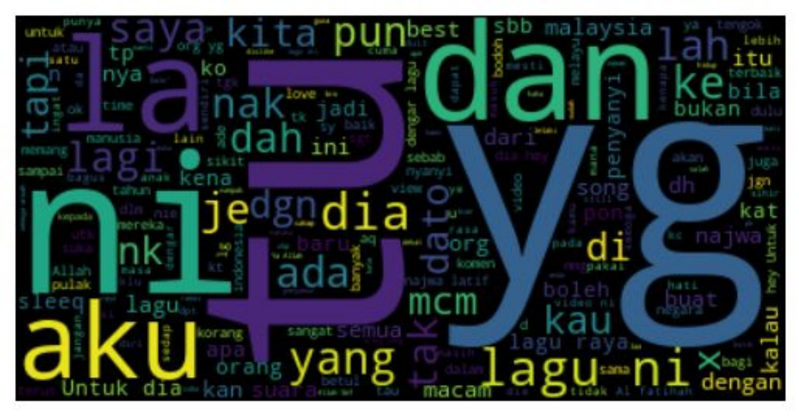
04

EDA

WORDCLOUD

There are many short forms, which can make understanding difficult, especially if there are multiple spellings per short form:

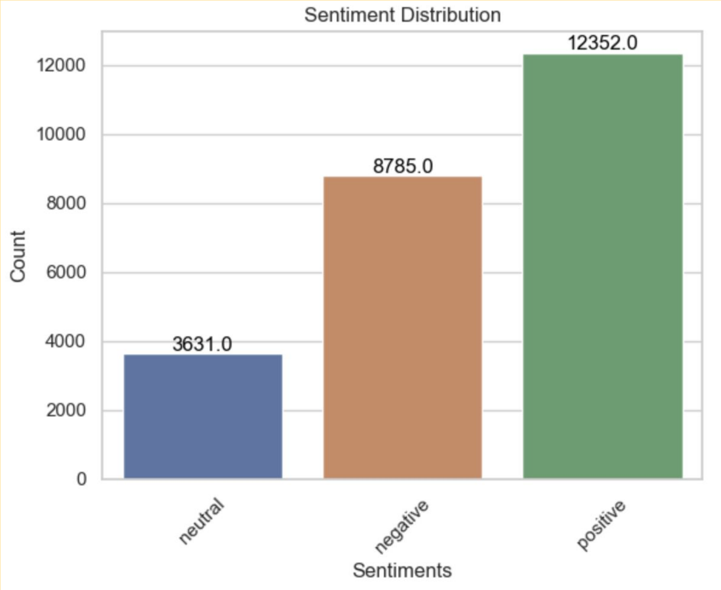
‘Yg’, ‘yang’ → ‘which’
‘Tu’, ‘itu’→ ‘that’
‘X’, ‘tak’, ‘tk’→ ‘not’
‘ni’ , ‘ini’ → ‘this’
‘Tp’, ‘tapi’ → ‘but’



Sentiment Distribution

Imbalanced classes - positive > negative > neutral

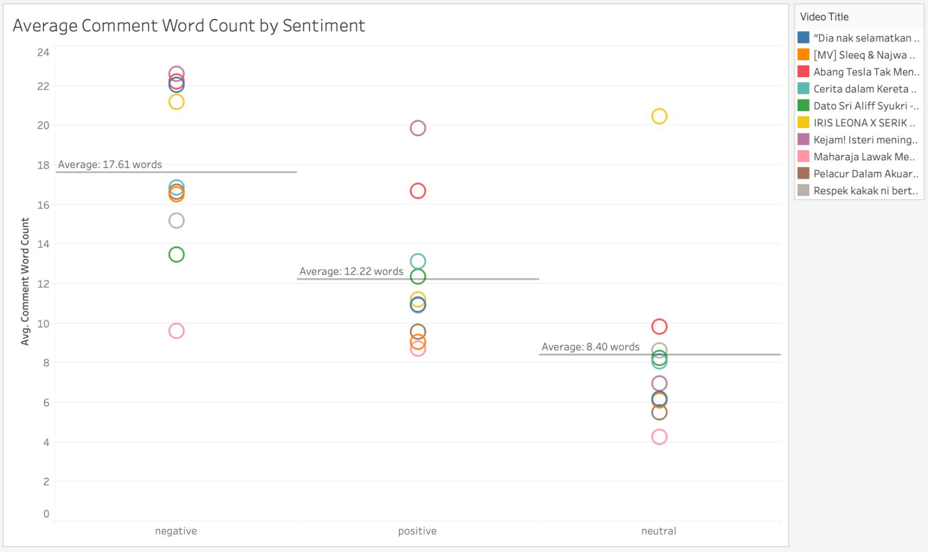
Need to take into account for modeling.



comment length

On average, negative comments tend to be longer than positive and neutral comments.

ChatGPT has a token limit, which means that longer comments might be truncated - or some comments might not be labelled

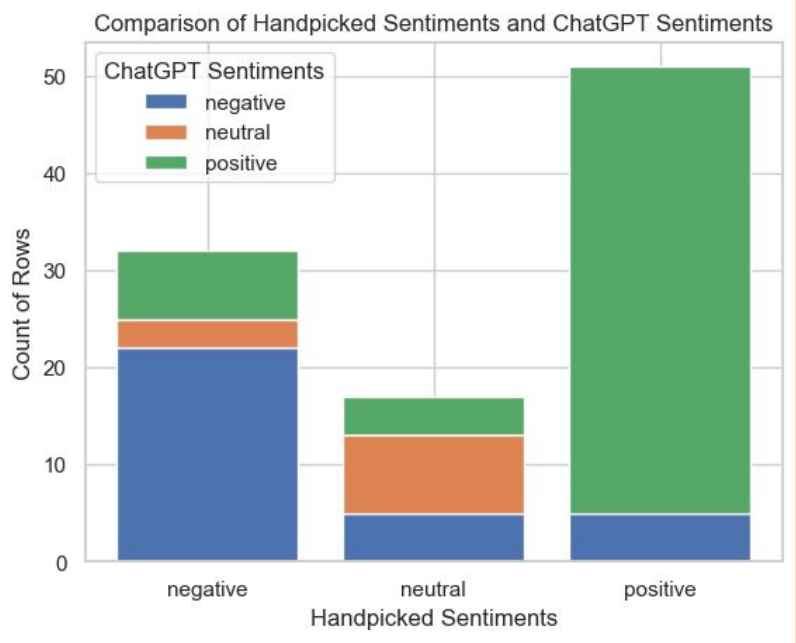

Average Comment Word Count by Sentiment

ChatGPT Performance

76% accuracy

Quite well balanced

Biggest challenge is the neutral sentiment

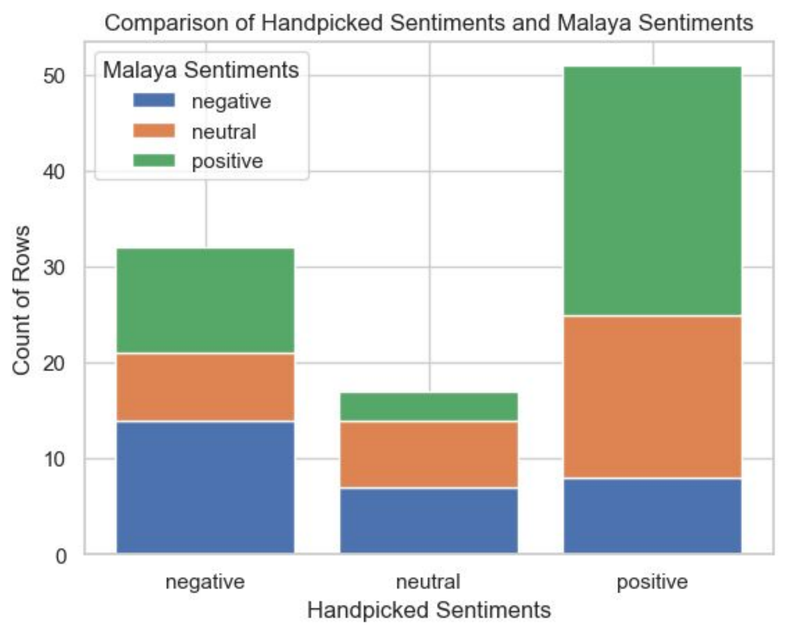


Malaya Performance

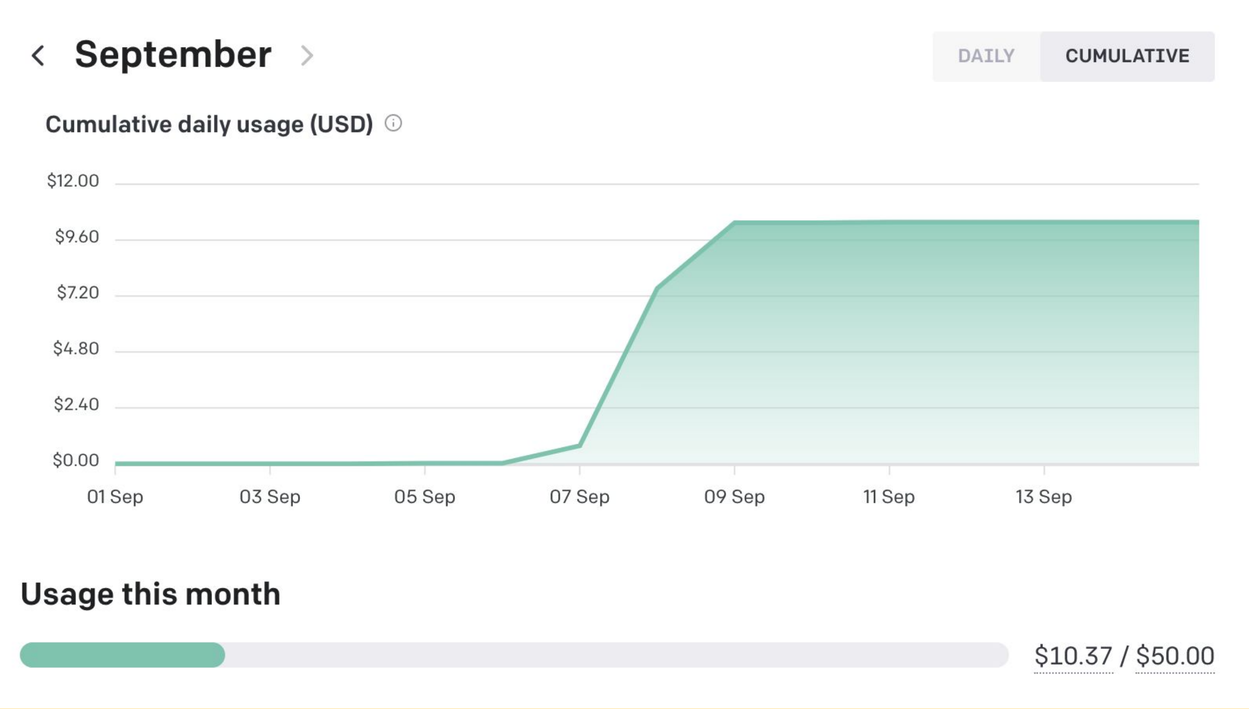
47% accuracy

Poor performance

Neutral comments tend to be considered to be positive, although the other two classes seem to be a little better



ChatGPT Usage



Cumulative daily usage (USD)

Usage this month

$10.37 / $50.00

DASHBOARD

https://public.tableau.com/views/SentimentAnalysisYouTubeComments/SentimentAnalysisDashboard

05

MODEL ANALYSIS

Best Model

We ran 8 models, and found that the best was:

Model 2: Multinomial Naive Bayes model + RandomOverSampler

	Train Set	Test Set
F1 Score	0.748	**0.701**

06

MODEL DEPLOYMENT

https://sentiment-analysis-bahasa-melayu.streamlit.app/

07

CONCLUSION

CONCLUSION

We managed to achieve a best-in-class sentiment analysis model for Malay social media texts.

LABELLING

ChatGPT is able to quite accurately label social media texts in Malay.

MODEL

Multinomial Naive Bayes + RandomOverSampler model ran the best in terms of test score and lack of overfitting

RESULTS

The model had a train and test F1 score of 0.748 and 0.701 respectively

Organisations and individuals can reliably use our tool to monitor their reputation among Malay speakers on social media.

Future works

Given more time, I would:

1 **Dive deeper into understanding intricacies of Malay social media text**

- Improved stopword generation
- Create Malay social media lexicon (similar to VADER)

2 **Get More Data**

- Improve sentiment analysis performance
- Create emotion analysis with more labelled data

THANK YOU

https://www.linkedin.com/in/ezracalis/

https://github.com/ezracalis

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**