TUNKU ABDUL RAHMAN UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

BMCS2123 NATURAL LANGUAGE PROCESSING

202309 Session, Year 2023/24

Assignment Documentation

Full Name: 1. Boon Yong Yeow 2. Ong Sheng Hao 3. Ong Ker Jing
Student ID: 1. 22PMR09689 2. 22PMR05894 3. 22PMR00748
Programme: Bachelors in Computer Science (Data Science)
Tutorial Class: G1
Project Title: Sentiment Analysis on Out-Of-Vocabulary (OOV) Malaysia Rojak Language
Module In-Charged:

Other team members' data

No	Student Name	Module In Charge
1	Boon Yong Yeow	XLM-T Transfer Learning (Fine-Tuning / Domain Adaptation)
2	Ong Sheng Hao	Data Preprocessing
3	Ong Ker Jing	Graphical User Interface

Lecturer: Dr Lim Khai Yin	Tutor : Dr Lim Khai Yin

1.0 Introduction

Sentiment analysis now encompasses procedures like sentiment opinion analysis, user opinion extraction, and text feature processing. Research currently in existence can be separated into two categories: machine learning-based sentiment analysis methods and traditional analysis methods based on sentiment lexicons. Each category has its own advantages: traditional analysis methods are characterised by their logical intelligence and interpretability, while machine learning methods are supported by data and statistics. In the big data era, machine learning techniques which typically consist of both supervised and unsupervised techniques have demonstrated their incredible performance (Gen Li et al., 2020). On the other hand, the critical performance of sentiment analysis can be significantly impacted by the presence of out of vocabulary (OOV), the words that are not included in a predetermined vocabulary (Yun Tang et al., 2020). Being a multilingual nation, Malaysia contains creoles and semicreoles that result from differing levels of proficiency in the various languages. Malaysian English (ME), a semi-creole form of English, is one such variation. ME that converges with local languages at all grammatical levels and combines into Malaysian Chinese is known as Bahasa Rojak (BR) (Ralf Vollmann & Soon Tek Wooi, 2019). For instance, "Jangan susah hati maa, lu punya bos mesti boleh kaw tim punya maa!", in this BR sentence, it included many words that are out of vocabulary such as “maa”, “li”, “kaw tim”. From the above example, we can know that BR consists of OOV to a certain extent.

Malaysians who are reacting on social media like to use BR. BR is difficult for any software or application to interpret and extract its true meaning (Abu Bakar et al., 2020). Standard text adheres to the language standard format, however noisy text does not, so pre-processing noisy and standard text is significantly different. Hence, performing sentiment analysis on OOV BR is a difficult task since it is hard to know the meaning of the words. The primary cause of the sentiment analysis model's poor performance was information loss brought on by the model representation's inability to handle OOV terms appropriately (Lochter, J. V., 2022). Because these BR contain many Out-Of-Vocabulary phrases, traditional sentiment analysis algorithms frequently produce incomplete and inaccurate sentiment assessments. Hence, there will be an issue when we preprocess the BR dataset, we need to figure out the ways to handle the OOV in order to handle the OOV to let the model understand its meaning. In order to sentiment the OOV, a diverse dataset of BR text that include both vocabulary and OOV words need to be collected. It will ensure that the dataset will be able to cover various domains and context to robust the sentiment analysis model. The most important part is data preprocessing, the core part is to handle OOV words by either replacing them with placeholder, removing them, or applying subword tokenization. Furthermore, using the dataset that were able to gather, the best language model needs to be selected to guarantee precise and significant outcomes. This is due to the fact that different language models perform differently when applied to various kinds of data; some are better with monolingual data, while others are better with multilingual data.

According to Bakar, M. F. R. Abu et al. (2020), the SA faced three primary difficulties with noisy Malay text. Noisy Malay texts are full of informal rules and patterns that are difficult to understand since they are often confusing and disorganised. Apart from that, noisy Malay text messages on social media platforms like Facebook and Twitter will occasionally change, making some of the guidelines or patterns outdated due to the constantly shifting nature of the trend. As a result, handling the noisy Malay text, which is constantly changing and getting more complex, is difficult. The absence of suitable resources for Malay South Africa is the next obstacle. The majority of SA has been completed for the English language, however there is still work to be done in the Malay language area. The final challenge is the limited available methods for dealing with Malay sarcasm. Sarcasm is occasionally used on social media platforms such as Facebook and Twitter. Therefore, it can be challenging to determine whether a sentence is sarcastic or not because its true meaning mostly depends on the desires of the individual.

The objective of the work is to develop a comprehensive lexicon. We aim to come up with a set of lexicons that help to handle the OOV occurring in the BR. This lexicon maps OOV words to either their full forms in their respective language or even directly to english. This helps preprocess the data to handle challenges such as misspellings, slang, abbreviations, and non-standard grammar commonly found in OOV languages. Aims to address the multilingual and cross-cultural aspects of sentiment analysis on OOV BR by considering the existing multilingual model and fine tuning it. Hence solve the problem of inability to interpret the meaning of OOV by a sentiment analysis model. In addition, the Sentiment Analysis (SA) model is constructed, making use of an extensive lexicon to accurately predict the sentiment associated with Bahasa Rojak inputs. A user-friendly interface has been carefully designed to enable users to quickly and easily determine the sentiment of the words they enter.

The importance of this research is that it addresses the existing problem of lack of technological Natural Language Processing (NLP) advancements in the domain of BR. By fine-tuning sentiment analysis models specifically for this unique language blend, it pioneers in filling a significant gap of the lack of such tools in the current existing market. Meanwhile it also serves as a guideline for interested researchers to understand and deal with intricacies that come with the handling of preprocessing and fine-tuning large language models with BR.

Sentiment analysis on Out-Of-Vocabulary (OOV) Bahasa Rojak is important because it can reveal complex feelings and viewpoints in this dynamic language environment. The analysis contributes to a better understanding of user attitudes and preferences in informal language use by accurately measuring sentiment in Bahasa Rojak. This understanding has many uses, including boosting consumer happiness, improving content recommendations, and strengthening strategies for interaction catered to the unique vocabulary of Bahasa Rojak (Adiba Nabiha et al., 2021).

2.0 Literature Review

2.1 Related work

2.1.1 XLM

Francesco Barbieri et al. (2022) presented a comprehensive framework for Twitter-based multilingual LMs, including the release of a new multilingual LM trained on almost 200M tweets to eliminate the reliance of current analysis on unclean pre-training and task-specific corpora as multilingual signals. Their focus is on sentiment analysis, collecting datasets in eight languages. After unifying and standardising the evaluation benchmark, Twitter-based multilingual LM is compared with a standard multilingual LM trained on general-domain corpora. The released model and code aim to support research in multilingual Twitter data (covering over thirty languages). Fine tuning is performed but there is some difference, instead of standard fine-tuning, adapter technique is integrated, by means of the LM is freezed and only fine-tunes one additional classification layer. There are speed and memory benefits to this strategy. Results indicate the effectiveness of the domain-specific language model, particularly in handling social media and multilingual sentiment analysis. Results indicate the effectiveness of the domain-specific language model, particularly in handling social media and multilingual sentiment analysis.

Nanda Putri Romadhona et al. (2020) built a corpus called BRCC to pre-train the language model of Bahasa Rojak and compiled SentiBahasaRojak, a dataset for sentiment analysis. Additionally, a brand-new pretrained model named Mixed XLM is proposed that maintains performance on monolingual data while simultaneously achieving the greatest results on code-mixing data in order to deal with the Bahasa Rojak code-mixing issue. Malay SMS rules are used to normalise short form words by adding some rules. Bahasa Rojak sentences in the bahasa rojak corpus are generated by the new data augmentation algorithm that is able to recognize 3 types of phrases in sentences and randomly select some of those for translation. The proposed Mixed XLM model can be used in downstream tasks containing code-mixing sentences, as long as it is pretrained on a code-mixing corpus. The MixedXLM model, pretrained on the Bahasa Rojak Corpus Collection (BRCC), demonstrates superior performance in sentiment analysis across three different language settings (English, Malay, and Bahasa Rojak), excelling in all evaluated domains after the evaluation.

2.1.2 Machine Learning Model

Ong Jun Ying et al. (2020) had developed a sentiment analysis model based on Convolutional Neural Network (CNN) architecture due to limited research on malay language. The process involved building a Word2Vec word embedding model to convert text input into numerical representation that is used to feed text data to CNN. After hyper-parameter tuning, the proposed CNN model is able to achieve an accuracy of 77.59% which surpasses similar work done on Bahasa Indonesia.

2.1.3 Lexicon and Dictionary

Keita Fujihira and Noriko Horibe (2020) proposed a multilingual sentiment analysis technique based on sentiment dictionary word-for-word translation to analyse the web user's opinion in various languages. The method of sentiment analysis is broken down into three stages. The first step is a morphological analysis of the text using the "TreeTagger" programme, which can analyse text in 25 languages, including German and English. After that, sentiments are extracted from words by using the "fastText" library to compare their similarity to a specific word, such as "love," and then translating the results using "Google Translate." Next, a sentiment dictionary in the original language is used to categorise the sentiment associated with these terms. In the last stage, the sentiment value of the entire text is calculated by combining the sentiments of each individual word into a three-dimensional vector that represents the feelings of Positive, Negative, and Neutral. In contrast to established classifiers such as "VADER" and "GCP," it showed advantages in terms of cost effectiveness by predicting sentiment values for every sentence translated word to word without requiring whole text translation. Evaluation experiments in the languages of Spanish, French, German, and English showed acceptable applicability for multilingual use.

Khalifa Chekima and Rayner Alfred (2018) constructed RojakLex1 lexicon which composite of 4 distinct lexicons, including MySentiDic: Malay lexicon, English Lexicon: English version of MySentiDic, Emoticon lexicon: list of commonly used emotion , and Neologism lexicon: made up of frequently occurring neologisms found in Malay social media text to tackle challenges faced by Malay social media text. For the method adopted, after collecting informal Malay text from the web, NLTK is used to tokenize collected data into sentences. In the text preprocessing part, Malay stopwords based on lists developed by Chekima,k and Alfred, R and English stopwords based on the existing Brown English Stopword list are removed. Afterward, sentences are mapped against RojakLex to identify polar words that exist in the data. This study also examines the impact of syntactical and grammatical rules on the perceived sentiment text. By recording 79.28%, the suggested method exhibits a significant improvement in accuracy over the baseline, which only recorded 51.38%.

2.1.4 Contextual Embedding

A unique Transformer-based word representation technique fused with Deep Intelligent Contextual Embedding ($DICE_T$) is proposed by Usman Naseem et al. (2020). Besides, an intelligent tweets pre-processor is designed to address noise in informal and unstructured tweets by rectifying spelling errors, performing sentiment-aware tokenization through the substitution of emoticons and slangs with actual words, and segmenting hashtags to enhance feature learning. The Viterbi algorithm is employed instead of the metaphone algorithm to determine the most likely sequence of POS tags for unseen sentences, allowing for efficient detection of the most probable tag sequence through dynamic programming. $DICE_T$ is capable of handling complicated word properties, word usage 660 in the noisy context of tweets, and other deep linkages. The results of the experiment demonstrate that, by using learning word representations, $DICE_T$ is also capable of handling ambiguities in language, such as polysemy, semantics, syntax, OOV terms, and sentiment knowledge. The findings indicate that the suggested sentiment analysis framework outperforms baseline approaches on the US airlines (+1.49%), airlines (+0.95%), and Emirates airlines (+1.80%) datasets. The findings from the experiment further validate $DICE_T$'s ability to handle linguistic difficulties and handle data analytics jobs involving imprecise and unclear data.

Jin Wang et al. (2020) present a contextual sentiment embeddings model that incorporates knowledge from labelled sentiment corpora and lexicons to solve the problem of traditional approaches ignoring the possibility that words might have various meanings in different contexts. . The model uses a two-layer GRU and was simultaneously trained on sentiment and semantic tasks. The suggested model outperforms previous suggested sentiment embeddings because it can handle uncommon or OOV emotional words using the WordPiece subword tokenizer. For example, the informal writing words coooool and wooooow can be tokenized into subword fragments, such as co, ##ooo, ##ol, and wo, ##ooo, ##ow, where the token is not at the front of the word indicated by the sign ##. Consequently, the emotional data from the ##ooo may be captured by the model. Empirical findings show that the proposed method outperforms previously proposed sentiment embeddings in sentiment classification by being able to distinguish between ambiguous meanings of the same word in various contexts.

2.2 Multilingual Models

David Vilares et al. (2017) conducted a study to tackle the problem of performing multilingual polarity classification on Twitter, comparing several techniques including multilingual model and monolingual model. The result of the experiment shows that a multilingual model is able to outperform the monolingual pipeline with language detection approximately 0.71% when the test dataset consists of multiple languages and english is not the majority language.

According to Bharathi Raja Chakravarthi et al. (2020), systems that are trained on monolingual data struggle with code-mixed data because of the complexity of language mixing at different levels. Dang Van Thin et al.'s (2023) experiment results further demonstrate that, in almost all datasets (Hotel Sentiment Analysis, VLSP Sentiment Analysis, Vietnamese Sentiment Analysis, UIT-VSFC, and Vietnamese Social Media Emotion Corpus, totaling 5 datasets), the multilingual XLM-R model performs better than all monolingual models except for PhoBERT. In certain datasets, the multilingual BERT outperformed the monolingual BERT as well.

From the above study, it can be concluded that the multilingual model is highly effective to be used in this study as the OOV bahasa rojak consists of multiple languages. Monolingual model seems not suitable for the multilingual dataset as mentioned above. Hence, a multilingual model is used for this study.

2.3 XLM-R

According to Deborah Aprilia Josephine et al. (2021), since labelled data is limited, transfer learning technique with full-fine tuning from an already existing model is utilised to recognise the texts. Multilingual-BERT-base-cased (mBERT) and XLM-RoBERTa-base (XLMR) as multilingual pretrained models with transformer architecture are used. With an average F1-score of 0.895, the evaluation result showed that the XLMR model outperforms mBERT for the entity extraction task.

Soumitra Ghosh et al. (2023) proposed proposed multi-task system is built on top of a cross-lingual embedding-based transformer model, XLMR, gain an accuracy of 71.61%, which yields better results than the existing XLMR-based state-of-the-art models on a similar task by 1.93%. An interesting finding is that the Multilingual BERT (mBERT) employed in the experiment underperforms when compared to both the XMLR baselines model and the suggested model by approximately 1%. It is said that mBERT displayed some cross-lingual characteristics, but it was not trained on cross-lingual data.

HuilingYou et al. (2021) had research on the application of three substantial language models for cross-lingual and multilingual word-in-context interpretation. The finding proved that it is better to fine-tune language models rather than using them as feature extractors. The results additionally show that XLMR outperformed mBERT in the cross-lingual context in terms of feature extraction and fine-tuning in 6 out of 9 dataset by average of 4.78%.

Gagan Bhatia et al. (2023) handle the task of sentiment analysis in 14 African languages by developing models for the zero-shot setting as well as models that are monolingual and multilingual in a fully supervised setting. Fine-tuning and further pretrained is performed on the models. The results indicate that Afro-XLMR-baseft outperforms other models in 5 languages with an average F1 score of 70.36. Besides that, the result shows that xlmr-base models outperform mbert-base models no matter in average or multilingual aspect.

According to Alexis Conneau (2020), by comparing against several cross-lingual benchmarks, the proposed dubbed XLM-R model performs noticeably better than multilingual BERT (mBERT). This comprises an average accuracy of +14.6% on XNLI, an average F1 score of +13% on MLQA, and an average F1 score of +2.4% on NER. When it comes to low-resource languages, XLM-R performs well, showing an improvement in XNLI accuracy of 15.7% for Swahili and 11.4% for Urdu when compared to previous XLM models.

Based on the previously mentioned study, it can be concluded that among multilingual models, the XLMR (Cross-lingual Language Model Representation) model performs better than the mBert (multilingual BERT) model. The XLMR model appears to outperform the mBert model in various aspects, emphasising its efficacy and effectiveness for tasks in the multilingual context. Hence, XLMR will be the model chosen for this study.

2.4 Hyperparameter Fine-Tuning

Previous research has established that fine-tuning of hyperparameters is able to help in improving the model accuracy. A study by Younas et al. (2020) explores sentiment analysis on code-mixed Roman Urdu-English social media text, employing multilingual models, involving XLM-R and mBERT. They fine-tune hyperparameters of both models on the batch size, learning rate and the number of epochs. The result shows that after fine-tuning of hyperparameters, XLM-R multilingual model performs much better than mBERT for code-mixed languages with significant improvement by 22% on F1 score. Similarly, Rønningstad (2023) explores sentiment analysis across 12 African languages using the XLM-T model, with hyperparameter adjustments including learning rate, warm up steps, weight decay, batch size, and epoch count. Their findings reveal that while fine-tuning enhances model performance for individual languages, it is less effective when trained on a dataset encompassing all languages. This suggests that hyperparameter tuning is crucial but must be tailored to the specific linguistic context to optimise sentiment analysis accuracy.

2.5 Ensemble Methods

Researchers have shown that ensemble methods provide better accuracy and efficiency. The study proposed by Al-Saqqa et al. (2018) employs an ensemble of machine learning classifiers approach, which incorporates the majority voting algorithm and four classifier algorithms, including Naive Bayes, Support Vector Machines, Decision Trees, and K-Nearest Neighbour. The aim of this study is to classify the sentiment polarity of Arabic text using ensemble methods. Based on the results, it appears that the ensemble voting method was effective and produced significantly better results with improvement of accuracy on average 10%, which is better than a single classifier.

A sentiment analysis on multiple domains has been proposed by Başarslan and Kayaalp (2023). They propose two ensemble learning methods, including Stacking and Majority Voting. The machine learning models include Naive Bayes, Support Vector Machines, Logistic Regression, and K-Nearest Neighbors. The use of machine learning models together in ensemble learning methods has demonstrated better results than using single classifiers. Majority Voting has shown to yield better results in all experiments when compared to Stacking with consistently more than average 0.3 on F1 score.

2.6 Chinese Word Segmentation

Segmentation is a crucial preprocessing step in Natural Language Processing, Due to the nature of the Chinese Language is highly context-dependent and ambiguous in terms of word representations, there has been various research conducted specifically on Chinese word segmentation in Natural Language Processing. Lei et al. (2021) research on Chinese segmentation in short text, uses five types of strategies of word segmentation, which involve Unigram, Bigram, Trigram Ltp, and Jieba. Shu et al. (2017) developed a char-based model that focuses on constructing feature functions using N-gram templates that contain information about the character as well as its subsequent characters. Li (2019) proposed both char-based and word-based models, and found that char-based models outperformed word-based models consistently.

Jieba is a Chinese language processing tool that is helpful in Chinese word segmentation. A study on lexicon-based Chinese Language sentiment analysis by Chen et al. (2019). As mentioned by the author, Jieba segmentation has many advantages. They mention that Jieba is easy to use since it can be easily plugged into Python code for sentiment analysis, and it is publicly accessible. On top of that, they also stated that Jieba will search for the most probable combination of words based on the maximum probability of paths. Also, it can process traditional Chinese word segmentation and support custom dictionaries in addition to three-word segmentation models (accurate, full, and search engine).

Li et al. (2018) researched sentiment analysis focusing on an emoticon polarity-aware recurrent neural network method. In the preprocessing stages of their study, they utilised the Jieba Python module for Chinese word segmentation to segment sentences from microblogs. The segmented text was then input into the Word2Vec model to train word vectors. This approach highlights the integral role of accurate word segmentation in preparing data for sentiment analysis and machine learning applications.

3.0 Proposed Methodology

3.1 Model Selection

3.1.1 Model Overview

The objective of this section is to provide an understanding of the two primary models considered for the task which are mBert (Multilingual BERT) and XLM-RoBERTa.

3.1.1.1 BERT and RoBERTa

Bidirectional Encoder Representation from Transformers (BERT) is a groundbreaking model in the field of NLP that changed the way machines understand human language. BERT uses 2 unsupervised tasks during pre-training: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM trains the model to predict randomly masked words in a sentence, providing deeper bidirectional context, while NSP trains BERT to understand the relationship between two sentences (Devlin et al., 2019).

Robustly Optimised BERT Pretraining Approach (RoBERTa) was introduced by Facebook AI in 2019 (Liu et al., 2019). It is an optimised version of BERT with modifications in the pre-training procedure that significantly improved performance. RoBERTa was trained on much larger dataset compared to its BERT counterpart. The NSP pretraining task was also eliminated after findings show that it does not contribute much to downstream task performance. RoBERTa only uses MLM as its pretraining task.

3.1.1.2 XLM-R

Cross-Lingual Language Model-RoBERTa (XLM-R) is the extension and improvement of the RoBERTa model tailored for cross-lingual understanding. XLM-R is trained on a significantly larger scale than its predecessors. It uses more than 2TB of filtered CommonCrawl Data.Unlike BERT which was trained on English text, or mBERT which was trained on Wikipedia text in multiple languages without the assurance of balanced and extensive coverage, XLM-R is trained on a more balanced multilingual dataset consisting of 100 languages including Malay (ms), English (en) and Chinese (zh) (Conneau et al., 2019).

3.1.1.3 XLM-T

XLM-T is a downstreamed application of XLM-R. It is a specialised domain adaptation of the XLM-R specifically fine-tuned for understanding and performing sentiment analysis tasks on twitter data. This Twitter-based multilingual model was trained on almost 200M tweets. XLM-T can often understand and perform tasks in languages it has never seen and been fine-tuned for due to the multilingual foundation of XLM-R. This ability is known as zero-shot learning.

3.2 Data Collection

In this research we utilised the BR Sentiment Analysis Dataset with labelled data from Putri et al. (2022) known as “SentiBahasaRojak”. To enhance generalizability of the sentiment analysis model, the dataset consisted of reviews from 3 separate domains being product reviews, movie reviews and stock reviews.

Both product and movie reviews originated from existing publicly available Malay datasets and were translated into BR using their self developed data augmentation method detailed in the original paper.

This approach involved a modified version of CoSDA-ML. The main difference is that their technique involved parsing sentences to identify phrases and selectively translating these into another language to ensure translated segments have inherent contextual meaning. This approach has been assessed using a Turing test format where 2 native Malay speakers were tasked to tell apart Augmented BR Data and Real-world BR Data from KLSE forum. Their results showed similar positive ratios in their Augmented Data and KLSE Data, which indicated the close similarity between the 2 data.

For stock reviews however, the data was collected from Malaysia's financial and stock market websites. 5 native speakers of BR were then hired to manually label these data. After the 5 experts were finished annotating all the posts, majority voting was carried out to determine the final labels for each and every one of the posts.

A total of 2285 observations were used for this study with 693, 699, and 893 posts for stocks, movies, and products respectively. The labels were binary with either -1 indicating negative sentiment or 1 indicating positive sentiment. Class imbalance was not a concern as the positive to negative class ratio was relatively balanced with 53% of the reviews being of negative sentiment and 47% of reviews being positive in sentiment.

3.3 Data Preprocessing

We have performed several preprocess steps, including:

1. **Remove URL:** URLs typically do not carry sentiment information. They are often used for linking to external content, and the sentiment analysis model may not derive meaningful sentiment from them. Removing URLs helps the model focus on the actual text content that contributes to sentiment.
2. **Convert the texts in datasets into lowercase:** Converting text to lowercase ensures that the same words are treated consistently. Without normalization, the model might treat words in uppercase and lowercase as different, potentially leading to the duplication of features and reduced generalization.
3. **Remove punctuations from the text except for hyphens and exclamation symbols:** In some cases, hyphens are used in compound words (e.g., "well-known," "high-level"). Removing hyphens in such cases may lead to misinterpretation or loss of meaning. For

instance, removing the hyphen in "well-known" would result in "wellknown," which could be a different word entirely. Exclamation symbols are often used to convey emphasis, excitement, or strong emotion in text. Removing them might alter the tone or sentiment of the text. For example, "Wow!" and "Wow" without the exclamation symbol can have different connotations.

4. **Remove Irregular space:** Irregular spaces, such as multiple consecutive spaces or leading/trailing spaces, can introduce inconsistencies in the representation of text. Removing irregular spaces ensures a consistent and standardized format, making it easier for the sentiment analysis model to process the data.
5. **Remove OOV words:** An OOV dictionary has been created by compiling information from both online[1] sources[2] and personal knowledge. This dictionary is utilized to transform out-of-vocabulary (OOV) words into a formal language, either in English or Malay. OOV words can cause errors in sentiment analysis predictions. If the model encounters a word it has never seen before, it may struggle to assign a sentiment score to that word, leading to potential inaccuracies in the overall sentiment prediction.
6. **Remove Stopwords:** English and Chinese stopwords are defined by the NLTK library's "stopwords" packages. However, for Malay stopwords, as the NLTK library doesn't offer specific packages for Malay stopwords, a list of Malay stopwords obtained from GitHub is utilized[3]. Stopwords are often very common in text but don't contribute much to the sentiment or meaning of a sentence. By removing them, the noise in the dataset can be removed, allowing the model to focus on more meaningful words that may convey sentiment.
7. **Remove Rare Words:** Rare words, especially those occurring only a few times, may not contribute significantly to the overall sentiment analysis task. Their presence increases the dimensionality of the data without providing substantial information. Removing them can lead to computational efficiency gains during training and inference.

[1] Zainal, M. K. (2016, January 19). *"Eh what does 'xtau' mean?" – A dictionary of Malay SMS short-forms for your sanity*. https://cilisos.my/bahasa-sms-shortforms-glossary/

[2] *Up-to-date list of Slangs for Text Preprocessing*. (n.d.). Kaggle.com. https://www.kaggle.com/code/nmaguette/up-to-date-list-of-slangs-for-text-preprocessing

[3] *Stopwords Malay (MS)*. (2021, May 5). GitHub. https://github.com/stopwords-iso/stopwords-ms/blob/master/stopwords-ms.txt

8. **Chinese word segmentation by using jieba:** Chinese sentences usually don't have space between the words, hence it is hard to tokenize. Consequently, Chinese segmentation is crucial for understanding the semantics of a sentence. Analyzing sentiments at the word level provides more granularity, allowing sentiment analysis models to capture the nuanced meanings associated with individual words. For example, the word “上海在中国” will be tokenize become “上海”, “在”, “中国” after chinese word segmentation. If without segmentation, then it is hard for the model to learn the meaning of the sentences.
9. **Convert label value from -1/1 to 0/1:** It is more common to use 0 and 1 for the binary classification evaluation.
10. **Remove instances:** Empty or null values in the 'text' column can introduce noise and inconsistency in the dataset. By removing instances with empty text entries, you ensure that the data is of higher quality and adheres to a consistent structure.
11. **Train test validation split:** For the train test validation split, the same method is applied for 3 datasets. First, 15% of the dataset is separated as the test dataset and the remaining 85% of the dataset is separated as the dataset1. After that, we split the dataset1 that contains 85% of the data and is further divided into 2 dataset with 17.65% of it becoming the validation set and remaining becoming the train set. The train, test and validation dataset for each dataset is then combined together.

3.4 Fine Tuning

This section documents the fine-tuning process of a multilingual sentiment analysis model based on the XLM-T architecture; with reference from the paper “XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond.” The paper’s objective was to downstream the pre-trained XLM-RoBERTa model to better understand and classify sentiment with the use of multilingual Twitter data (Francesco Barbieri et al., 2022).

Our aim is to adapt and extend this approach and further fine-tune XLM-T specifically for Bahasa Rojak using SA dataset from Putri et al. (2022). We seek to enhance the model’s ability to accurately capture and interpret the nuanced sentiment expressed in BR thereby contributing to the broader field of sentiment analysis in mixed-language contexts.

3.4.1 Hardware and Software

The fine-tuning utilised Google Colaboratory (Colab) which is an accessible cloud service that provides a Python notebook running in a Virtual Machine. This study leveraged several Using an NVIDIA Tesla T4 GPU. As for software, this research utilised the Transformers library by Hugging Face, which provides a comprehensive suite of pre-trained models, including XLM-T. Additionally for managing the training process, Version 2.1.2 of PyTorch, a machine learning library was utilised as the backend framework for the Transformers library.

3.4.2 Hyperparameters

This research builds on the fine-tuning starter code provided by the authors to facilitate the use of their XLM-T framework. The starter code provided a detailed fine-tuning script including some crucial predefined parameters like Learning Rate, Epochs, Batch Size, Warm Up Steps, Weight Decay and etc. Due to computing resources limitations of google colab, the Batch Size hyperparameter used for this study (16) did not reflect the original values provided (32). Other than that the hyperparameter used remained the same:

1. Learning Rate (LR=2e-5): Using a conservative learning rate of 2e-5 specifically for fine tuning purposes, mitigating the risks of pre-trained model to not deviate too drastically from originally learned patterns (Kandel & Castelli, 2020).

2. Epochs (EPOCHS =1): Conservative approach of using only one epoch to prevent overfitting (Xu et al., 2023).

3. Batch Size (BATCH_SIZE=16): Originally 32, a batch size of 16 was selected to ensure balance between the model's learning stability and computing efficiency during training (Memory Constraint in Google Colabs).

4. Maximum Training Examples (MAX_TRAINING_EXAMPLES=-1): This setting indicated that the entire training dataset was to be used for fine-tuning, ensuring comprehensive exposure to the language and sentiment characteristics of Bahasa Rojak.

5. Warmup Steps (warmup_steps=100): The gradual increase to specified learning rate to ensure stability in model training early on.

6. Weight Decay (weight_decay=0.01): Serving a regularisation technique, a 0.01 weight decay helps prevent overfitting and the overall model remains generalizable (Liu et al., 2020).

Some of the other hyperparameters included were Logging Steps (10), Save and Eval Steps (10), and Best Model Loading (True).

3.4.3 Modelling and Evaluation

The model was trained using the Hugging Face 'Trainer' API built on top of the PyTorch framework which is one of the most widely-used deep learning libraries. For evaluation, standard classification metrics such as precision, recall and F1-score were employed to provide a comprehensive understanding of the model's capabilities in correctly identifying sentiments (Umarani et al., 2021).

4.0 Experiment and Results

This section delves into a series of experiments carried out to find the best method of fine-tuning the XLM-T model for SA of Bahasa Rojak. Given the memory constraints of Google Colab and the challenges associated with extensive hyperparameter tuning, the focus shifted towards experimenting with different combinations of preprocessing techniques to improve SA results (HaCohen-Kerner et al., 2020).

4.1 Preprocessing Approaches

A total of 4 preprocessing approaches were employed on SentiBahasaRojak to find best input data that produces optimal results for the SA model. This produced 4 Fine-Tuned XLM-T Models trained on different training data. The specific preprocessing processes are shown below in table 1 and figure 1 provides a graphical representation of model development flow.

PA1: Serves as the baseline. The preprocessing steps included removing irrelevant information to Bahasa Rojak like Removing URLs, standardising text like conversion to lowercase and handling OOV, and noise removal like removing stop words and punctuations. The aim was to clean and normalise text while retaining its essential sentiments.

PA2: An extension of V1, V2 introduces a process called Chinese Character Segmentation using the Jieba package. Jieba uses a probability-based approach that calculates the most probable combination of words. This method helps in accurately identifying word boundaries in Chinese where there is the absence of space making segmentation a challenging task Chen et al. (2019).

PA3: An extension of V2, V3 introduces another process to remove rare words that only occur once in the dataset. This step was taken in hopes that reducing noise and lexical diversity in the training data would lead to a model focusing on impactful language features.

PA4: A more less aggressive approach to preprocessing training data. This pipeline focuses on only removing URLs, replacing irregular spacing and handling OOV. Stopwords, punctuations and uppercase were maintained. Findings show that stemming and stopword removal has potential risk to result in the decrease of accuracy in sentiment analysis tasks in Bahasa Indonesia (Pradana & Hayaty, 2019; HaCohen-Kerner et al., 2020).

Table 1: List of Preprocessing Steps for V1 - V4

Preprocessing Method	Model 1 (V1)	Model 2 (V2)	Model 3 (V3)	Model 4 (V4)
Remove URLs	✔	✔	✔	✔
Convert Lowercase	✔	✔	✔	-
Remove Punctuations	✔	✔	✔	-
Remove Irregular Spaces	✔	✔	✔	✔
Handle OOV	✔	✔	✔	✔
Remove Stopwords	✔	✔	✔	-
Chinese Character Segmentation	-	✔	✔	-
Remove Rare Words	-	-	✔	-

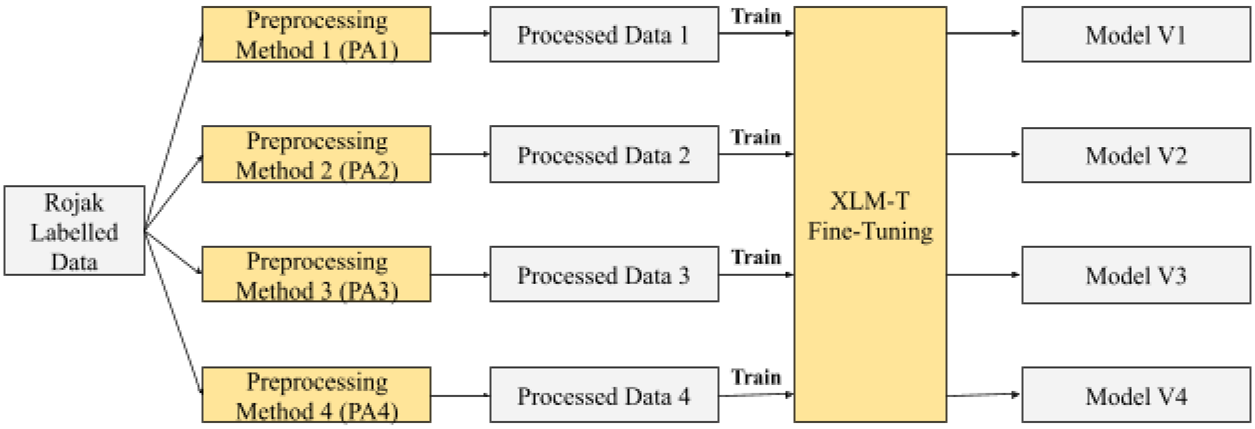


Figure 1: Diagram Illustrating Experiment with Different Preprocessing Approaches

4.2 Model Results

Table 2 below shows the classification report depicting a detailed comparison across Model V1 to V4, evaluating their performance on BR test data. Model V1 demonstrates the highest recall value for class 0 (Negative Sentiments) with a score of 0.84 indicating its effectiveness in finding all negative sentiments in the dataset. Model V2 shows a more balanced performance in correctly identifying both negative and positive sentiments. Model V3, while presenting the highest precision for the negative sentiment class, falls short in recall for negative sentiments identification. Model V4 outshines all the other models across almost all metrics, with the highest performance in precision, recall (Positive Sentiments), F1-Scores and highest accuracy of 0.845. While Model V4 fine-tuned with PA4 managed the most reliable performance the strength of Model V1 can be exploited to ensure that no negative sentiments are missed.

Table 2: Model 1 (V1) to Model 4 (V4) Classification Report

	Precision		**Recall**		**F1-Score**		**Accuracy**
	0	1	0	1	0	1	
Model V1	0.716	0.830	**0.840**	0.702	0.773	0.760	0.767
Model V2	0.768	0.771	0.735	0.801	0.751	0.786	0.770
Model V3	0.794	0.703	0.691	0.802	0.739	0.749	0.744
Model V4	**0.861**	**0.833**	0.802	**0.884**	**0.831**	**0.858**	**0.845**

4.3 Ensemble Approach

To ensure a robust and adaptable approach to performing sentiment analysis on the nuanced nature of BR. We leveraged an ensemble majority approach that allows us to capitalise on these diverse strengths (Al-Saqqa et al., 2018; Başarslan and Kayaalp, 2023).

Initial experiment explored the feasibility of pairing models (V1+V4, V2+V4, and V3+V4) for an ensemble majority voting approach in sentiment analysis on Bahasa Rojak. The objective was to examine the frequency of contradiction between the predictions between each pair. Unsurprisingly, a significant number of instances were omitted from final scoring due to such contradictions illustrated in Table 3 below. The finding exposed the limitations of deploying a straightforward ensemble method and underscore the necessity of a different strategy that can resolve contradictions between model predictions.

Table 3: Experimenting with Paired Models for Ensemble Approach

	Accuracy	Support	Omitted	Total Test Data
V1 + V4	0.877	138	**205**	343
V2 + V4	0.869	137	**206**	343
V3 + V4	0.713	80	**263**	343

As a result of the high number of omissions observed in the initial pair-based ensemble approach, we decided to introduce the Original XLM-T model as a tie-breaker in cases of contradictory predictions. This strategy aims to leverage the robustness of the Original XLM-T model, which has been extensively trained on diverse Twitter data. Additionally, we conducted experiments using Preprocessing Approaches 1 to 4 to determine which preprocessing pipeline produces data that the Original XLM-T model performs best on. This step is crucial for optimising performance and ensuring that our ensemble method not only resolves disagreements more effectively but also capitalises on the strengths of each individual model involved. As shown in Table 4, the original XLM-T model performs optimally on PA4.

Table 4: Original XLM-T Model Classification Report using Different Preprocessing Approaches

	Precision		**Recall**		**F1-Score**		**Accuracy**
	0	1	0	1	0	1	
PA1	0.807	0.883	0.889	0.798	0.846	0.839	0.842
PA2	0.807	0.883	0.889	0.798	0.846	0.839	0.842
PA3	0.853	0.842	0.853	0.842	0.853	0.842	0.848
PA4	**0.857**	**0.918**	**0.912**	**0.866**	**0.884**	**0.891**	**0.888**

After incorporating the Original XLM-T model into the ensemble methodology, a systematic testing of various combinations of the models was carried out to determine which ensemble would yield the best results. The goal was to identify the combination that not only maximised accuracy but also provided a balanced representation of sentiment classes. As a result, illustrated in table 5 below, ensemble combination V1+V4+V5 (Original XLM-T) emerged as the combination that showcased best performance in all performance metrics in the classification report.

Table 5: Ensemble Combinations Classification Report

	Precision		**Recall**		**F1-Score**		**Accuracy**
	0	1	0	1	0	1	
V1 + V4 + V5	**0.815**	**0.844**	**0.825**	**0.835**	**0.820**	**0.840**	**0.830**
V2 + V4 + V5	0.800	0.792	0.750	**0.835**	0.774	0.813	0.795
V3 + V4 + V5	0.583	0.500	0.462	0.620	0.515	0.554	0.535

After thorough experimentation to obtain a sentiment analysis model for BR, three primary models were identified as competitive: Model V4, Original XLM-T, and the ensemble combination of V1+V4+V5. The metric achieved by these 3 models are show in Table 6 below.

Table 6: Comparison between Competitive BR SA Models

	Precision		**Recall**		**F1-Score**		**Accuracy**
	0	1	0	1	0	1	
XLMT (PA4)	0.857	**0.918**	**0.912**	0.866	**0.884**	**0.891**	**0.888**
Model V4	**0.861**	0.833	0.802	**0.884**	0.831	0.858	0.845
V1 + V4 + V5	0.815	0.844	0.825	0.835	0.820	0.840	0.830

5.0 Graphical User Interface

The Graphical User Interface (GUI) of our sentiment analysis tool is divided into three sections including Homepage, Text Prediction, and Web Scraping (see Appendix for GUI display). It was designed to offer a user-friendly experience and empowers users to gain a holistic view of sentiment trends, providing both detailed insights and a big-picture understanding of the sentiments expressed in their text or Twitter data. Figure 2 showcases the flow of how sentiment analysis final predictions were made, illustrating the step-by-step process from user input to final decision.

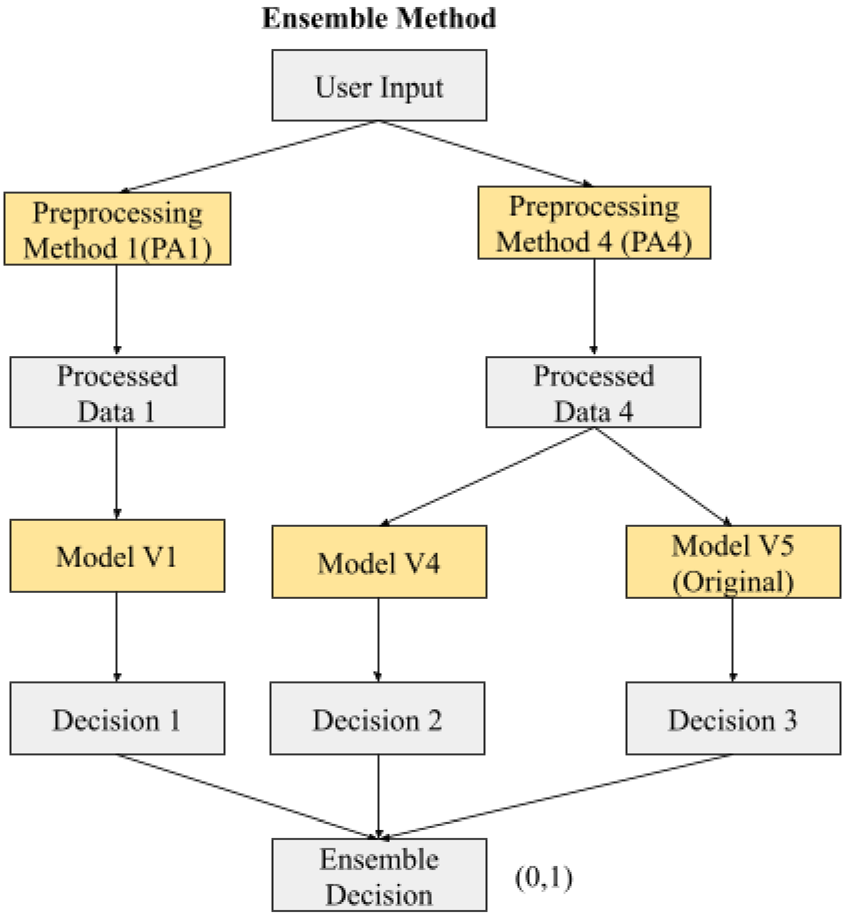


Figure 2: Sentiment Analysis Workflow Diagram

5.1 Homepage

The Homepage serves as the gateway to the sentiment analysis tool. It provides a brief introduction to the project, outlining the purpose and capabilities of this tool. Additionally, clear user instructions are provided, ensuring that users can easily navigate and utilise the tool effectively from their initial interaction.

5.2 Text Prediction

In the Text Prediction section, users are invited to input text for sentiment analysis. Initially, the text is summarised using a pre-trained Bidirectional and Auto-Regressive Transformers (BART) model (Lewis, 2019), which is renowned for its effectiveness in generating clear and meaningful summaries. This model distils the text into a condensed version, ensuring that users can quickly grasp the essence of their input. The next part displays abstract topics derived from the Latent Dirichlet Allocation (LDA) model presented by Blei et al. (2003), which is a useful tool for Topic Modelling, helping to highlight the main themes. Finally, the Sentiment Analysis results present predictions from models V1 and V4, indicating negative or positive sentiment, as well as sentiment scores from model V5. An ensemble method that combines the outcomes of models V1, V4, and V5 is then employed to reach a final decision, providing comprehensive sentiment analysis.

5.3 Web Scraping

The Web Scraping section extends the tool's functionality by allowing users to analyse sentiment for specific hashtags or user accounts on Twitter. Users can specify their query, choose the number of tweets to analyse, and select the type of tweets to fetch. The results are then displayed in a user-friendly format, showcasing a list of scraped posts, interactive tweet statistics, sentiment proportions, a dynamic sentiment over time graph, and illustrative word clouds for both positive and negative sentiments.

6.0 Discussion and Conclusion

Discuss the outputs of your work. Any limitations/ shortcomings or advantages? Besides, it is good to include recent literature (as discussed in the earlier section) for comparing your work to highlight your work in view of recent development and challenges in the field. Finally, conclude your work.

6.1 Challenges in Bahasa Rojak Sentiment Analysis

This study addresses a crucial gap in sentiment analysis for Bahasa Rojak, a language characterised by its multilingual nature and common occurrences of out-of-vocabulary (OOV) words. The study illustrates that complex, informal rules and evolving patterns of BR present significant challenges for traditional sentiment analysis methods even after fine-tuning downstreamed LLMs (XLM-T).

6.2 Importance of Preprocessing

The research highlights the importance of preprocessing textual data that is often unstructured to handle challenges of BR. Various preprocessing approaches (PA1 to PA4) were employed, each building on top of one another to experiment with the aim to find the optimal scale of preprocessing that produces the best results. Through this experimentation, it was evident that preprocessing significantly impacts the performance and accuracy of SA models as performance decreased with the introduction of more preprocessing steps like Chinese Word Segmentation and Removal of Rare Words. On the contrary, SA Model V4 which was fine-tuned using minimal preprocessing out of PA1-PA4 outperformed the rest of the models in almost every other performance metric used. This showcased that potential contextual loss occurred with added preprocessing procedures. This is aligned with the findings of Pradana and Hayaty (2019) which stated SA model accuracy tends to decrease with the implementation of stemming and removal of stop words.

6.3 Ensemble Approach

The exploration into SA models for BR resulted in identifying three primary models as competitive contenders: model V4, Original XLM-T, and the ensemble combination of V1+V4+V5. Notably, the Original XLM-T, even without further tuning, outperformed its counterparts, achieving an overall accuracy of 0.888. Its high precision rates for both positive (0.918) and negative sentiments (0.912) suggest a strong ability to understand and classify the nuanced sentiments within BR. The superior performance of untuned original XLM-T models underscore several potential limitations of the study like insufficient data, limited computing resources and specific preprocessing approaches resulting in the loss of contextual information and the ability to generalise.

A detailed comparison between Model V4 and the ensemble combination (V1+V4+V5) reveals insightful shifts in performance. The ensemble method's decreased precision for negative sentiments (from 0.861 to 0.815) alongside an increased recall (from 0.802 to 0.825) suggests a strategic trade-off, favouring the broader capture of negative sentiments over pinpoint accuracy. Conversely, its performance on positive sentiments showed a slight improvement in precision but a decrease in recall. This nuanced shift highlights the ensemble method's tendency to lean more towards identifying negative sentiments, a potentially desirable feature in contexts where missing out on negative sentiments could be more detrimental.

6.4 Limitations and Future Work

The main limitations of this study are mainly the constrained computing resources available from google colaboratory cloud services. This restricted the extent of model training and experimentation. The daily usage limits on computing resources prolonged the experimentation process. This constraint not only affected the speed at which the models could be iteratively improved but also limited the scope for hyperparameter tuning. The EPOCH and BATCH SIZE were crucial parameters that were subject to these limitations, necessitating a compromise between the number of epochs and the size of the batches to fit within the available resources.

Consequently, this may have impacted the thoroughness of the model training and the ultimate performance of the sentiment analysis models.

The other limitation is the scarcity of labelled BR SA datasets. This study primarily relied on SentiBahasaRojak datasets from Putri et al. (2022). That was constructed using their own data augmentation algorithm to convert originally SA datasets in Malay to BR. This might not best reflect the nuances and variety of real-life BR used in Malaysian society.This might be reflected in the models ability to generalise to real life unstructured text data.

Future work in this should focus on expanding the scope and diversity of used Bahasa Rojak datasets. Using a more extensive and varied collection of real-life BR text for training and fine-tuning purposes might significantly enhance models' ability to understand the intricacies of BR. Additionally, more advanced model architectures could be further exploited potentially leading to better results. Future studies should benefit from sufficient computational resources, allowing for a more customizable hyperparameter configuration thus optimising model performance. Finally, incorporating user feedback could also be explored and provide an avenue for reinforcement learning techniques where the SA model could improve by learning from real-time user interactions and feedback.

6.5 Conclusion

The research undertaken in this study represents a step forward in the sentiment analysis of Bahasa Rojak (BR). By employing various preprocessing approaches and fine-tuning advanced multilingual models based on XLM-R architecture, the study has shed light on the complexities and intricacies inherent in understanding and interpreting the sentiments expressed in this unique and dynamic language. Overall this research found a total of viable 3 models that are able to perform sentiment analysis tasks with reasonable accuracy in the multilingual context of Bahasa Rojak. These models, namely Model V4, Original XLM-T, and the ensemble combination model of V1+V4+V5, each having their own unique strengths and with each achieving an accuracy of 0.83, 0.845, and 0.888 respectively.

7.0 References

Al-Saqqa, S., Obeid, N., & Awajan, A. (2018). Sentiment Analysis for Arabic Text using Ensemble Learning. IEEE Xplore. https://doi.org/10.1109/aiccsa.2018.8612804

Bakar, M. F. R. Abu, Idris, N., Shuib, L., & Khamis, N. (2020). Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work. *IEEE Access*, 8, 24687-24696. doi: https://doi.org/10.1109/ACCESS.2020.2968955

Barbieri, F., Espinosa-Anke, L., & José Camacho-Collados. (2021). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2104.12250

Başarslan, M. S., & Kayaalp, F. (2023). Sentiment analysis with ensemble and machine learning methods in multi-domain datasets. Turkish Journal of Engineering, 7(2), 141–148. https://doi.org/10.31127/tuje.1079698

Bhatia, G., Adebara, I., Elmadany, A., & Abdul-Mageed, M. (2023). UBC-DLNLP at SemEval-2023 Task 12: Impact of Transfer Learning on African Sentiment Analysis. In AfriSenti 2023 @ ACL 2023. arXiv:2304.11256 [cs.CL]. https://doi.org/10.48550/arXiv.2304.11256

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. https://doi.org/10.5555/944919.944937

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) 2020. arXiv:2006.00210 [cs.CL]. https://doi.org/10.48550/arXiv.2006.00210

Chekima, K., & Alfred, R. (2018). Sentiment Analysis of Malay Social Media Text. Lecture Notes in Electrical Engineering, 205–219. https://doi.org/10.1007/978-981-10-8276-4_20

Chen, J., Becken, S., & Stantić, B. (2019). Lexicon based Chinese language sentiment analysis method. Computer Science and Information Systems, 16(2), 639–655. https://doi.org/10.2298/csis181015013c

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Guillaume Wenzek, Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised Cross-lingual Representation Learning at Scale. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1911.02116

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020 (+ updated results). arXiv:1911.02116 [cs.CL]. https://doi.org/10.48550/arXiv.1911.02116

Deborah Aprilia Josephine, Ayu Purwarianti, & Ekaputra, F. J. (2021). *Knowledge Graph Construction using Information Extraction of Indonesia Cosmetic Product Text in Bahasa Indonesia*. https://doi.org/10.1109/icaicta53211.2021.9640251

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/pdf/1810.04805.pdf

Fujihira, K., & Horibe, N. (2020, September 1). Multilingual Sentiment Analysis for Web Text Based on Word to Word Translation. IEEE Xplore. https://doi.org/10.1109/IIAI-AAI50415.2020.00025

Ghosh, S., Priyankar, A., Ekbal, A., & Bhattacharyya, P. (2022). Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. *Knowledge-Based Systems*, 110182. https://doi.org/10.1016/j.knosys.2022.110182

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. PLOS ONE, 15(5), e0232525. https://doi.org/10.1371/journal.pone.0232525

Kandel, I., & Castelli, M. (2020). How Deeply to Fine-Tune a Convolutional Neural Network: A Case Study Using a Histopathology Dataset. Applied Sciences, 10(10), 3359. https://doi.org/10.3390/app10103359

Lei, Q., Li, H., & Chen, Y. (2021). How Does Chinese Segmentation Strategy Effect on Sentiment Analysis of Short Text? IEEE Access. https://doi.org/10.1109/prml52754.2021.9520738

Lewis, M. (2019, October 29). BART: Denoising Sequence-to-Sequence Pre-training for natural language generation, Translation, and Comprehension. arXiv.org. https://arxiv.org/abs/1910.13461

Li, D., Rzepka, R., Ptaszyński, M., & Araki, K. (2018). Emoticon-Aware Recurrent Neural Network Model for Chinese Sentiment Analysis. IEEE Conference Publication | IEEE Xplore. https://doi.org/10.1109/icawst.2018.8517232

Li, G., Zheng, Q., Zhang, L., Guo, S., & Niu, L. (2020). Sentiment Infomation based model for Chinese text sentiment analysis. *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*. https://doi.org/10.1109/auteee50969.2020.9315668

Li, X. (2019, May 14). Is word segmentation necessary for deep learning of Chinese representations? arXiv.org. https://arxiv.org/abs/1905.05526

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July 26). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv.org. https://arxiv.org/abs/1907.11692

Liu, Z., Cui, Y., & Chan, A. B. (2020). Improve Generalization and Robustness of Neural Networks via Weight Scale Shifting Invariant Regularizations. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2008.02965

Lochter, J. V., Silva, R. M., & Almeida, T. A. (2022). Multi-level out-of-vocabulary words handling approach. *Knowledge-Based Systems*, *251*, 108911. https://doi.org/10.1016/j.knosys.2022.108911

Nabiha, A., Mutalib, S., & Malik, A. M. (2021). Sentiment analysis for informal Malay text in Social Commerce. *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. https://doi.org/10.1109/aidas53897.2021.9574436

Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis. *Future Generation Computer Systems*, *113*, 58–69. https://doi.org/10.1016/j.future.2020.06.050

Pradana, & Hayaty. (2019). The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-language Texts. Kinetik, 4(4), 375–380. https://doi.org/10.22219/kinetik.v4i4.912

Putri, N., Lu, S.-E., Lu, B.-H., Tzong, R., & Tsai, H. (2022). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset (pp. 4418–4428). https://aclanthology.org/2022.coling-1.389.pdf

Romadhona, N. P., Lu, S.-E., Lu, B.-H., & Tsai, R. T.-H. (2022, October 1). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. ACLWeb; International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.389/

Romadhona, N. P., Lu, S.-E., Lu, B.-H., & Tsai, R. T.-H. (2022). BRCCandSentiBahasaRojak: The First BahasaRojak Corpus for Pretraining and Sentiment Analysis Dataset. Proceedings of the 29th International Conference on Computational Linguistics, 4418–4428. https://aclanthology.org/2022.coling-1.389.pdf

Rønningstad, E. (2023, April 27). UIO at SemEval-2023 Task 12: Multilingual fine-tuning for sentiment classification in low-resource languages. arXiv.org. https://arxiv.org/abs/2304.14189

Shu, X., Wang, J., Shen, X., & Qu, A. (2017). Word segmentation in Chinese language processing. Statistics and Its Interface, 10(2), 165–173. https://doi.org/10.4310/sii.2017.v10.n2.a1

Tang, Y., Tang, C., & Zhu, C. (2020). Resolve out of vocabulary with long short-term memory networks for morphology. *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. https://doi.org/10.1109/icaica50127.2020.9182586

Umarani, V., Julian, A., & Deepa, J. (2021). Sentiment Analysis using various Machine Learning and Deep Learning Techniques. Journal of the Nigerian Society of Physical Sciences, 385–394. https://doi.org/10.46481/jnsps.2021.308

Van Thin, D., Hao, D. N., & Nguyen, N. L.-T. (2023). Vietnamese Sentiment Analysis: An Overview and Comparative Study of Fine-tuning Pretrained Language Models. *ACM Transactions on Asian and Low-Resource Language Information Processing*. https://doi.org/10.1145/3589131

Vilares, D., Alonso, M. A., & Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, *53*(3), 595–607. https://doi.org/10.1016/j.ipm.2017.01.004

Vollmann, R., & Wooi, S. (2019). The sociolinguistic status of Malaysian English. Grazer Linguistische Studien, 91, 133-150. https://doi.org/10.25364/04.46:2019.91.5

Wang, J., Zhang, Y., & Yu, L.-C. (2022). Contextual sentiment embeddings via bi-directional GRU language model. *Knowledge-Based Systems*, *235*, 107663–107663. https://doi.org/10.1016/j.knosys.2021.107663

Xu, C., Coen-Pirani, P., & Jiang, X. (2023). Empirical Study of Overfitting in Deep Learning for Predicting Breast Cancer Metastasis. Cancers, 15(7), 1969–1969. https://doi.org/10.3390/cancers15071969

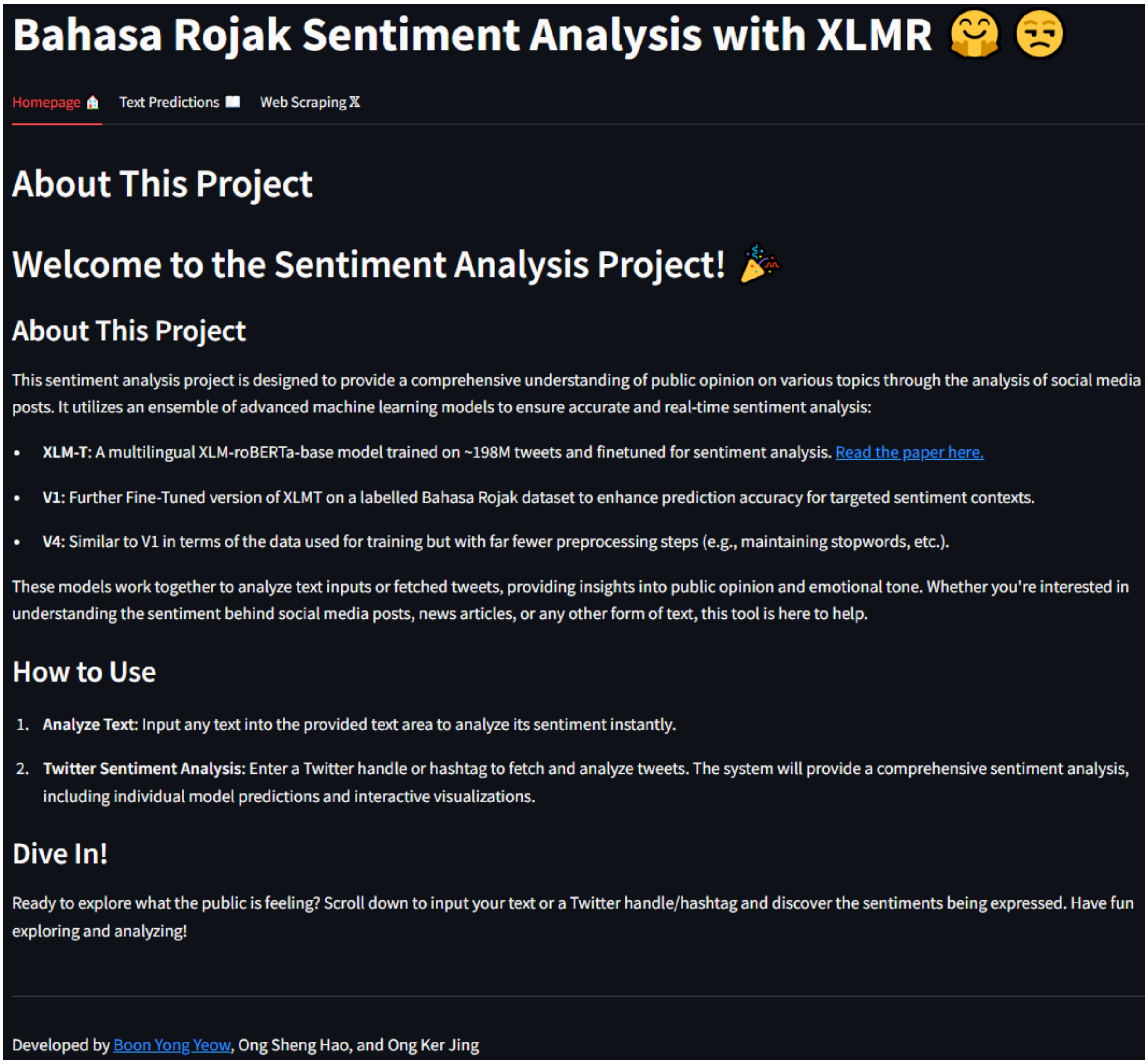
Ying, O. J., Zabidi, M. M. A., Ramli, N., & Sheikh, U. U. (2020). Sentiment analysis of informal Malay tweets with deep learning. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *9*(2), 212. https://doi.org/10.11591/ijai.v9.i2.pp212-220

You, H., Zhu, X., & Stymne, S. (2021). Uppsala NLP at SemEval-2021 Task 2: Multilingual Language Models for Fine-tuning and Feature Extraction in Word-in-Context Disambiguation. To appear at SemEval-2021. arXiv:2104.03767 [cs.CL]. https://doi.org/10.48550/arXiv.2104.03767

Younas, A., Nasim, R., Ali, S., Wang, G., & Qi, F. (2020). Sentiment Analysis of Code-Mixed Roman Urdu-English Social Media Text using Deep Learning Approaches. IEEE Conference Publication | IEEE Xplore. https://doi.org/10.1109/cse50738.2020.00017

8.0 Appendix

Below is the Graphical User Interface (GUI) design of our sentiment analysis tool. These screenshots provide a visual representation of the tool's layout, functionalities, and user interaction points.



Bahasa Rojak Sentiment Analysis with XLMR 🤗 😒

About This Project

Welcome to the Sentiment Analysis Project! 🎉

About This Project

This sentiment analysis project is designed to provide a comprehensive understanding of public opinion on various topics through the analysis of social media posts. It utilizes an ensemble of advanced machine learning models to ensure accurate and real-time sentiment analysis:

- **XLM-T:** A multilingual XLM-roBERTa-base model trained on ~198M tweets and finetuned for sentiment analysis. Read the paper here.
- **V1:** Further Fine-Tuned version of XLMT on a labelled Bahasa Rojak dataset to enhance prediction accuracy for targeted sentiment contexts.
- **V4:** Similar to V1 in terms of the data used for training but with far fewer preprocessing steps (e.g., maintaining stopwords, etc.).

These models work together to analyze text inputs or fetched tweets, providing insights into public opinion and emotional tone. Whether you're interested in understanding the sentiment behind social media posts, news articles, or any other form of text, this tool is here to help.

How to Use

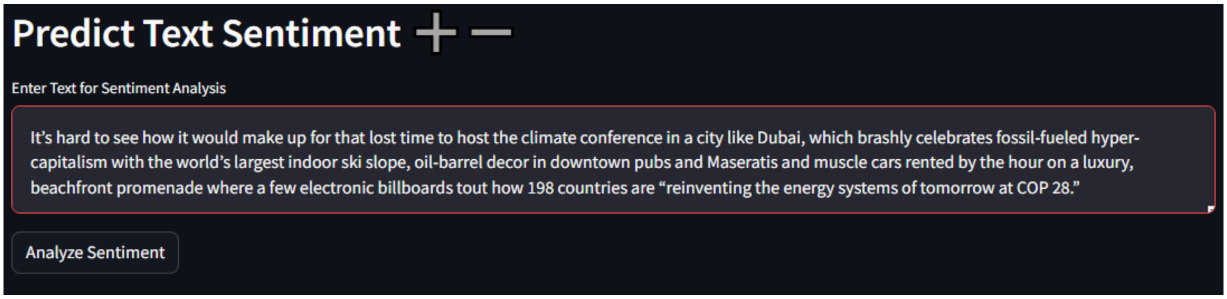
1. **Analyze Text:** Input any text into the provided text area to analyze its sentiment instantly.
2. **Twitter Sentiment Analysis:** Enter a Twitter handle or hashtag to fetch and analyze tweets. The system will provide a comprehensive sentiment analysis, including individual model predictions and interactive visualizations.

Dive In!

Ready to explore what the public is feeling? Scroll down to input your text or a Twitter handle/hashtag and discover the sentiments being expressed. Have fun exploring and analyzing!

Developed by Boon Yong Yeow, Ong Sheng Hao, and Ong Ker Jing

Figure 3: Homepage

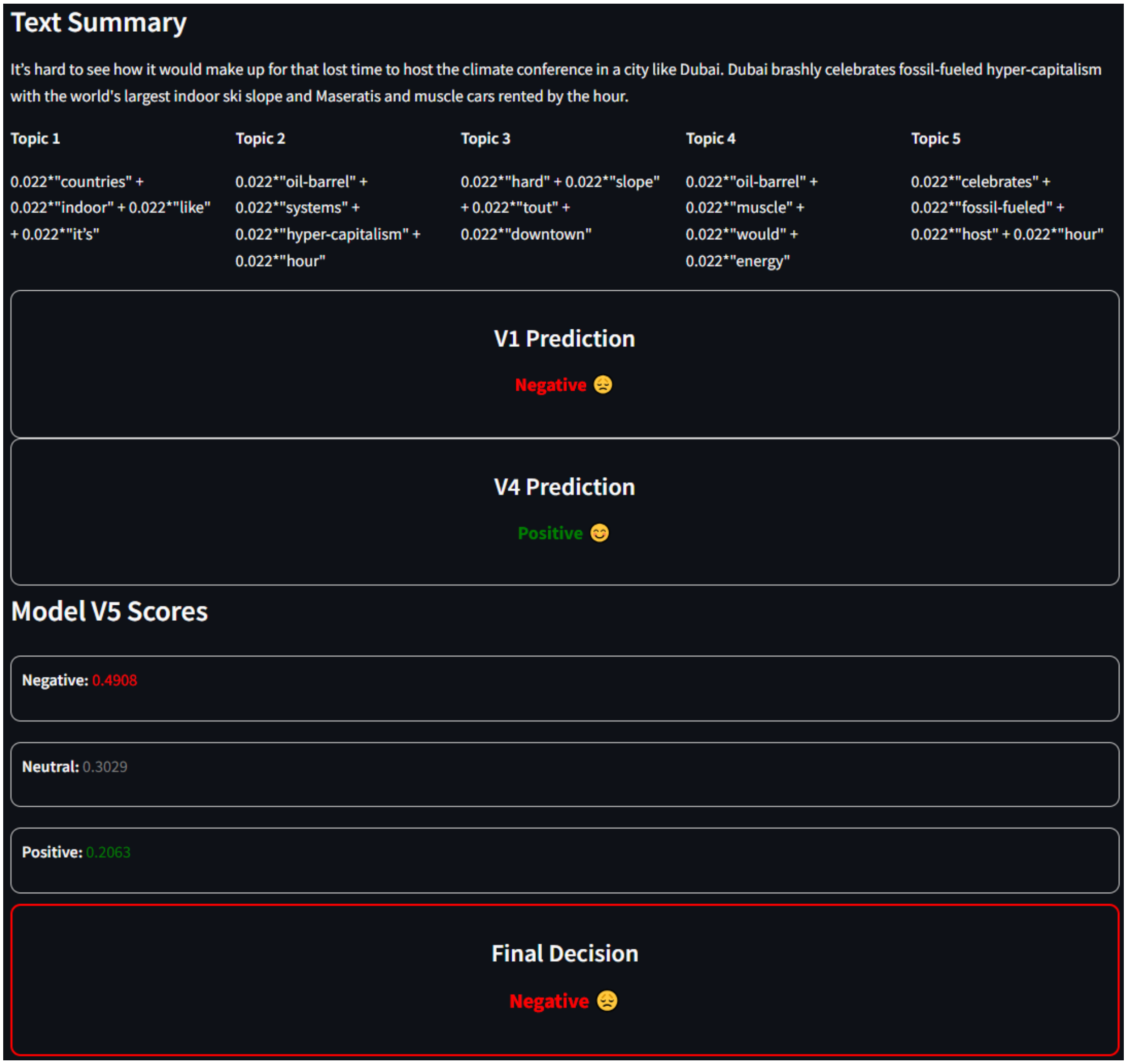


Predict Text Sentiment

Enter Text for Sentiment Analysis

It’s hard to see how it would make up for that lost time to host the climate conference in a city like Dubai, which brashly celebrates fossil-fueled hyper-capitalism with the world’s largest indoor ski slope, oil-barrel decor in downtown pubs and Maseratis and muscle cars rented by the hour on a luxury, beachfront promenade where a few electronic billboards tout how 198 countries are “reinventing the energy systems of tomorrow at COP 28.”

Figure 4: Text Predictions Sections



Text Summary

It’s hard to see how it would make up for that lost time to host the climate conference in a city like Dubai. Dubai brashly celebrates fossil-fueled hyper-capitalism with the world's largest indoor ski slope and Maseratis and muscle cars rented by the hour.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0.022*"countries" + 0.022*"indoor" + 0.022*"like" + 0.022*"it's"	0.022*"oil-barrel" + 0.022*"systems" + 0.022*"hyper-capitalism" + 0.022*"hour"	0.022*"hard" + 0.022*"slope" + 0.022*"tout" + 0.022*"downtown"	0.022*"oil-barrel" + 0.022*"muscle" + 0.022*"would" + 0.022*"energy"	0.022*"celebrates" + 0.022*"fossil-fueled" + 0.022*"host" + 0.022*"hour"

V1 Prediction

Negative 😞

V4 Prediction

Positive 😊

Model V5 Scores

Negative: 0.4908

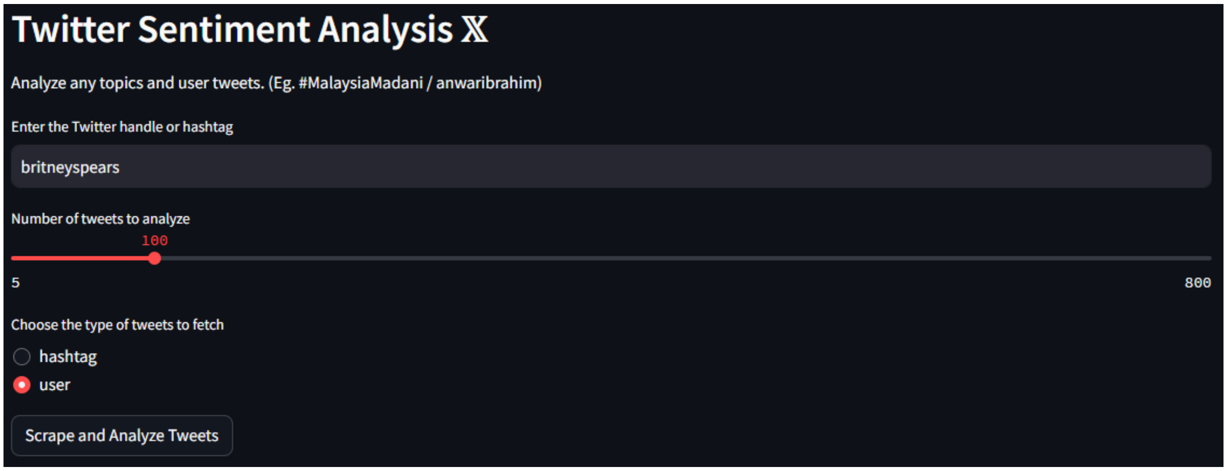
Neutral: 0.3029

Positive: 0.2063

Final Decision

Negative 😞

Figure 5: Result Display of Text Predictions



Twitter Sentiment Analysis 𝕏

Analyze any topics and user tweets. (Eg. #MalaysiaMadani / anwaribrahim)

Enter the Twitter handle or hashtag

britneyspears

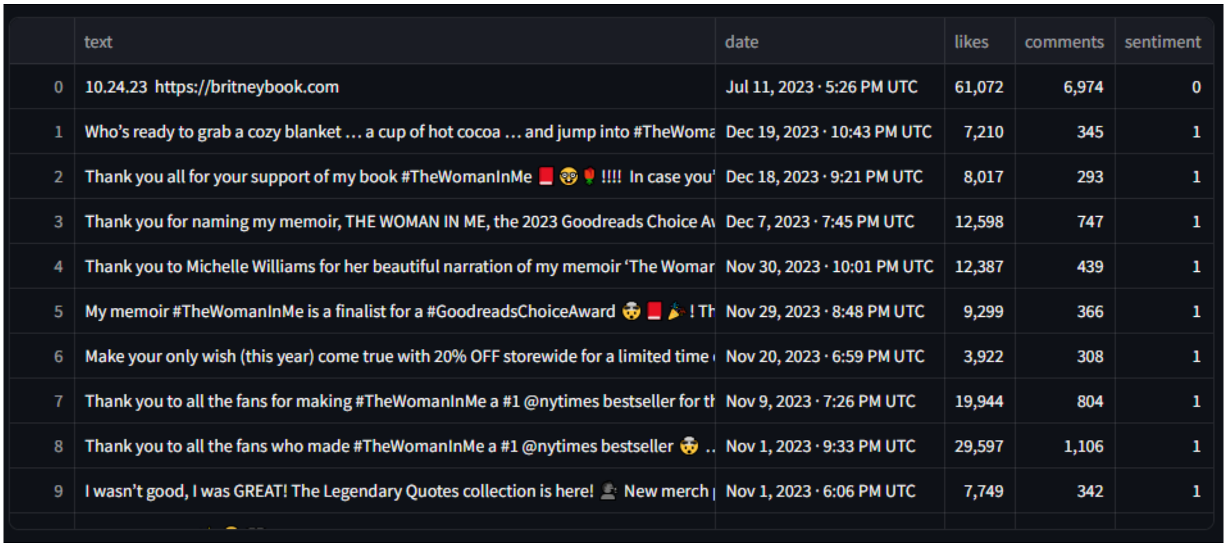
Number of tweets to analyze

100

Choose the type of tweets to fetch

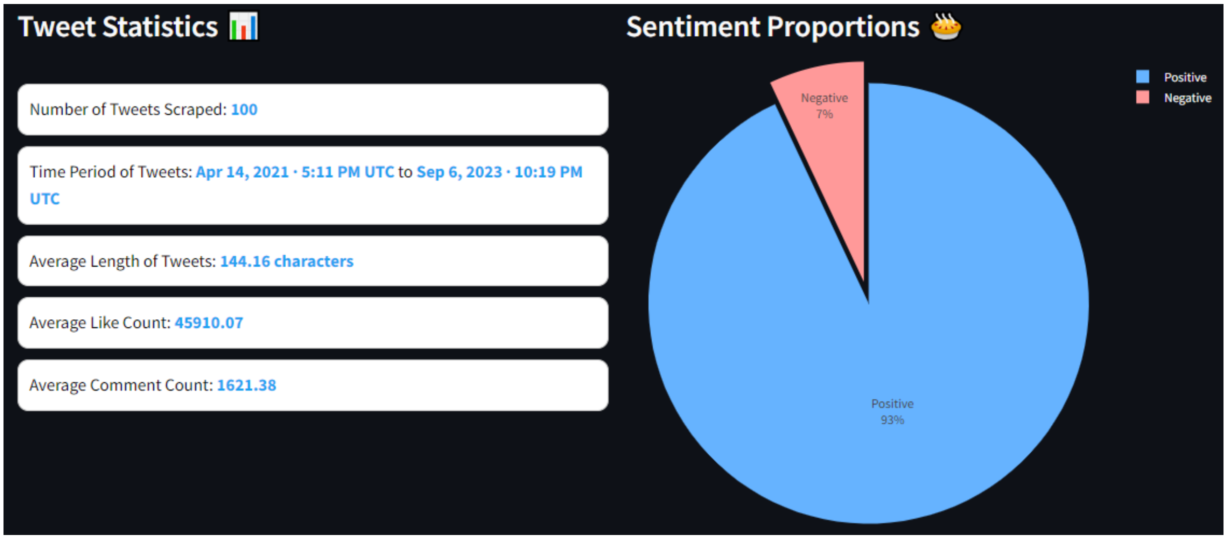
○ hashtag
● user

Figure 6: Web Scraping Section



	text	date	likes	comments	sentiment
0	10.24.23 https://britneybook.com	Jul 11, 2023 · 5:26 PM UTC	61,072	6,974	0
1	Who's ready to grab a cozy blanket … a cup of hot cocoa … and jump into #TheWoma	Dec 19, 2023 · 10:43 PM UTC	7,210	345	1
2	Thank you all for your support of my book #TheWomanInMe 📕 🤓 🌹 !!!! In case you'	Dec 18, 2023 · 9:21 PM UTC	8,017	293	1
3	Thank you for naming my memoir, THE WOMAN IN ME, the 2023 Goodreads Choice Av	Dec 7, 2023 · 7:45 PM UTC	12,598	747	1
4	Thank you to Michelle Williams for her beautiful narration of my memoir 'The Womar	Nov 30, 2023 · 10:01 PM UTC	12,387	439	1
5	My memoir #TheWomanInMe is a finalist for a #GoodreadsChoiceAward 🤯 📕 🎉 ! Th	Nov 29, 2023 · 8:48 PM UTC	9,299	366	1
6	Make your only wish (this year) come true with 20% OFF storewide for a limited time	Nov 20, 2023 · 6:59 PM UTC	3,922	308	1
7	Thank you to all the fans for making #TheWomanInMe a #1 @nytimes bestseller for th	Nov 9, 2023 · 7:26 PM UTC	19,944	804	1
8	Thank you to all the fans who made #TheWomanInMe a #1 @nytimes bestseller 🤯 ..	Nov 1, 2023 · 9:33 PM UTC	29,597	1,106	1
9	I wasn't good, I was GREAT! The Legendary Quotes collection is here! 👩‍⚖️ New merch	Nov 1, 2023 · 6:06 PM UTC	7,749	342	1

Figure 7: List of Scraped Post



Tweet Statistics 📊

Number of Tweets Scraped: 100

Time Period of Tweets: Apr 14, 2021 · 5:11 PM UTC to Sep 6, 2023 · 10:19 PM UTC

Average Length of Tweets: 144.16 characters

Average Like Count: 45910.07

Average Comment Count: 1621.38

Sentiment Proportions 🥧

Figure 8: Tweet Statistics and Sentiment Proportions

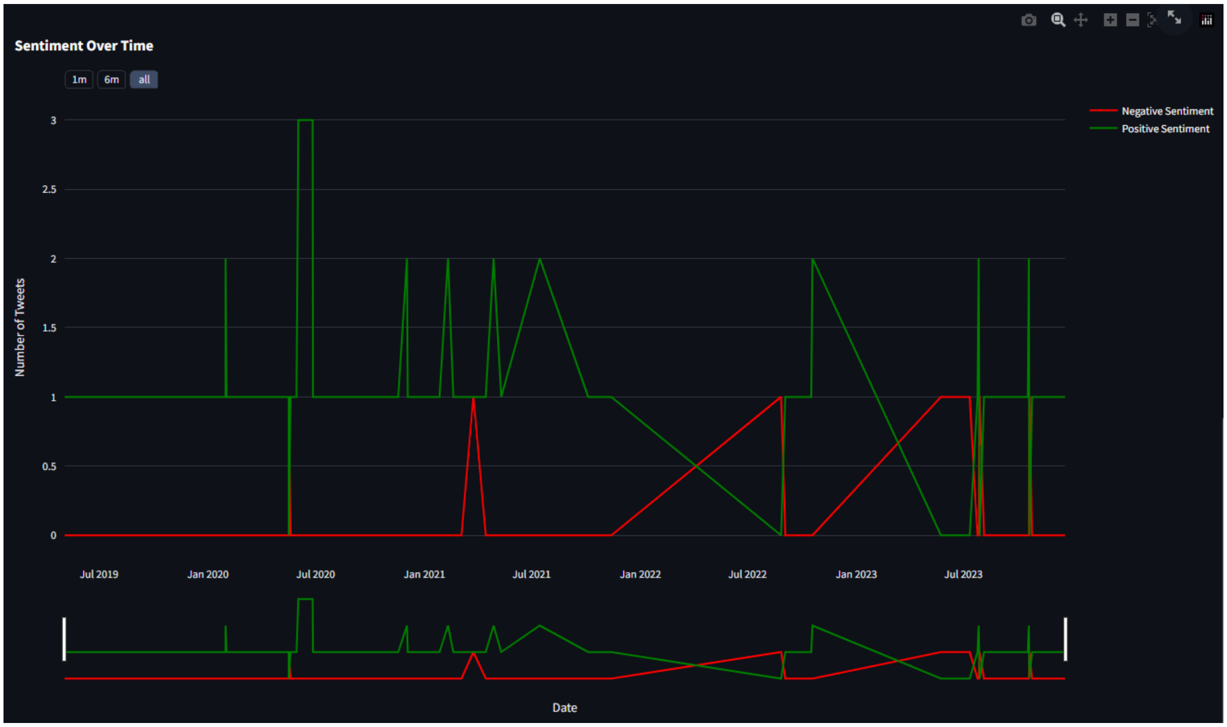
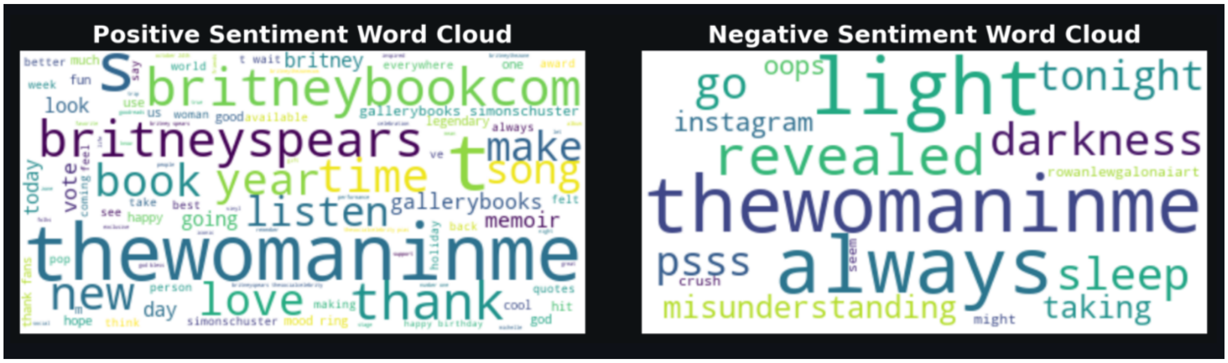


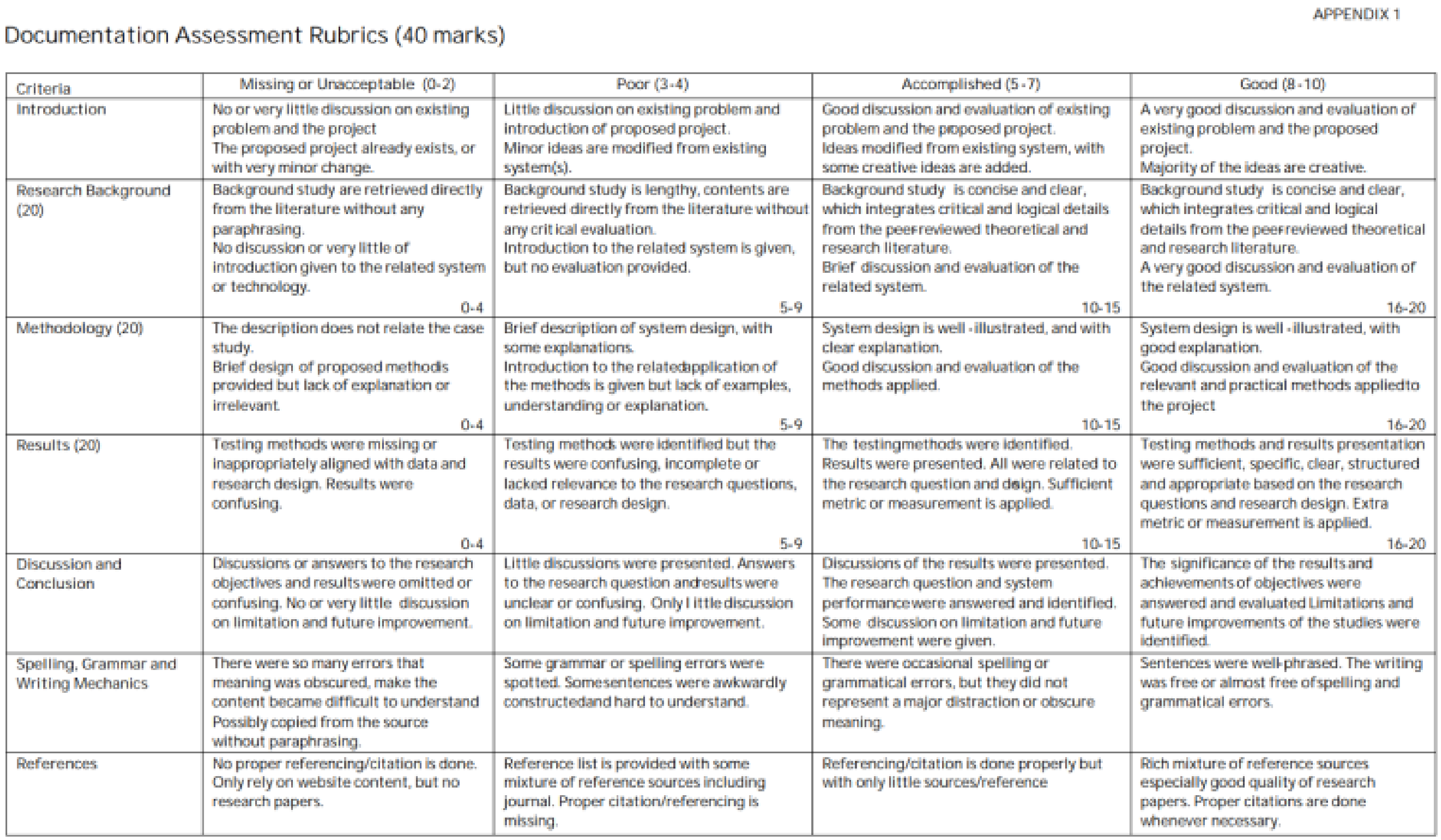
Figure 9: Sentiment Over Time Graph


Positive Sentiment Word Cloud

Negative Sentiment Word Cloud

Figure 10: Positive and Negative Word Cloud

Marking Rubric



APPENDIX 1

Documentation Assessment Rubrics (40 marks)

Criteria	Missing or Unacceptable (0-2)	Poor (3-4)	Accomplished (5-7)	Good (8-10)
Introduction	No or very little discussion on existing problem and the project The proposed project already exists, or with very minor change.	Little discussion on existing problem and introduction of proposed project. Minor ideas are modified from existing system(s).	Good discussion and evaluation of existing problem and the proposed project. Ideas modified from existing system, with some creative ideas are added.	A very good discussion and evaluation of existing problem and the proposed project. Majority of the ideas are creative.
Research Background (20)	Background study are retrieved directly from the literature without any paraphrasing. No discussion or very little of introduction given to the related system or technology. 0-4	Background study is lengthy, contents are retrieved directly from the literature without any crit ical evaluation. Introduction to the related system is given, but no evaluation provided. 5-9	Background study is concise and clear, which integrates critical and logical details from the peer-reviewed theoretical and research literature. Brief discussion and evaluation of the related system. 10-15	Background study is concise and clear, which integrates critical and logical details from the peer-reviewed theoretical and research literature. A very good discussion and evaluation of the related system. 16-20
Methodology (20)	The description does not relate the case study. Brief design of proposed methodis provided but lack of explanation or irrelevant. 0-4	Brief description of system design, with some explanations. Introduction to the relatedapplication of the methods is given but lack of examples, understanding or explanation. 5-9	System design is well -illustrated, and with clear explanation. Good discussion and evaluation of the methods applied. 10-15	System design is well -illustrated, with good explanation. Good discussion and evaluation of the relevant and practical methods appliedto the project 16-20
Results (20)	Testing methods were missing or inappropriately aligned with data and research design. Results were confusing. 0-4	Testing methods were identified but the results were confusing, incomplete or lacked relevance to the research questions, data, or research design. 5-9	The testingmethods were identified. Results were presented. All were related to the research question and design. Sufficient metric or measurement is applied. 10-15	Testing methods and results presentation were sufficient, specific, clear, structured and appropriate based on the research questions and research design. Extra metric or measurement is applied. 16-20
Discussion and Conclusion	Discussions or answers to the research objectives and resultswere omitted or confusing. No or very little discussion on limitation and future improvement.	Little discussions were presented. Answers to the research question andresults were unclear or confusing. Only l ittle discussion on limitation and future improvement.	Discussions of the results were presented. The research question and system performancewere answered and identified. Some discussion on limitation and future improvement were given.	The significance of the resultsand achievementsof objectives were answered and evaluated Limitations and future improvements of the studies were identified.
Spelling, Grammar and Writing Mechanics	There were so many errors that meaning was obscured, make the content became difficult to understand Possibly copied from the source without paraphrasing.	Some grammar or spelling errors were spotted. Somesentences were awkwardly constructedand hard to understand.	There were occasional spelling or grammatical errors, but they did not represent a major distraction or obscure meaning.	Sentences were well-phrased. The writing was free or almost free ofspelling and grammatical errors.
References	No proper referencing/citation is done. Only rely on website content, but no research papers.	Reference list is provided with some mixture of reference sources including journal. Proper citation/referencing is missing.	Referencing/citation is done properly but with only little sources/reference	Rich mixture of reference sources especially good quality of research papers. Proper citations are done whenever necessary.

APPENDIX 2

Prototype Assessment Rubrics (60 marks)

No	Item	Criteria		
		Poor	Accomplished	Good
1	U ser interface / output (20)	Poor or confusing design of UI or output, which provides inadequate information/outputs Most of the information/outputs generated are less accurate. Layout of information is not organized. 0-8	A dequate information/outputs needed are generated The information/output generated are accurate but some with errors. Layout of information is organized. 9-15	A ll the necessary information/outputs are generated All or most of the information/outputs generated are accurate. Minor errors can be ignored. Layout of information is w ell-organized. 16-20
2	Programming (30)	The end product fails with many logic errors, many actions lacked exception handling. Solutions are over-simplified. Programming skill needs improvement. Minimal validations are provided. Business rules are not validated 0-10	Major parts are logical, but some steps to complete a specific job may be tedious or unnecessarily complicated. Program algorithm demonstrates acceptable level of complexity. The student is qualified to be a programmer. Important and necessary validations are provided 11-19	Correct and logical flow, exceptions are handled well. Demonstrates appropriate or high level of complex algorithms and programming skills. Thorough and thoughtful validations are provided. All important business rules are validated 20-30
3	Degree of completion (20)	Too much still remain to be done. Basic requirements are not fulfilled. The end product produces enormous errors, faults or incorrect results 0-8	All required features present in the interface within the required scope, but some are simplified. Or one or two features are missing. The system is able to run with minor errors 9-15	All required features present in the interface within or beyond the required scope no bugs during demonstration 16-20
4	System implementation (20)	The end product is produced with different system design or approach, which is not related to the initial proposal 0-8	The end product conforms to most of the system design, but some are different from the specification 9-15	The end product fully conforms to the proposed system design 16-20
5	On-the-spot coding (10)	The student is unclear about the work produced, sometimes not even knowing where to find the source code. 0-4	The student knows the code whereabouts, but sometimes may not be clear why the work was done in such a way. 5-7	The student is clear about every piece of the work done. 8-10